Preprint typeset in JHEP style - HYPER VERSION

Michaelmas Term, 2019

Cosmology

University of Cambridge Part II Mathematical Tripos

David Tong

Department of Applied Mathematics and Theoretical Physics,
Centre for Mathematical Sciences,
Wilberforce Road,
Cambridge, CB3 OBA, UK

http://www.damtp.cam.ac.uk/user/tong/cosmo.html
d.tong@damtp.cam.ac.uk

Recommended Books and Resources

Cosmology textbooks sit in one of two camps. The introductory books do a good job of describing the expanding universe, but tend to be less detailed on the hot Big Bang and structure formation. Meanwhile, advanced books which cover these topics assume prior exposure to both general relativity and statistical mechanics. This course sits somewhere between the two.

The first two books below cover the material at an elementary level; the last three are more advanced.

- Barbara Ryden, *Introduction to Cosmology*

A clearly written book that presents an excellent, gentle introduction to the expanding universe, with subsequent chapters on thermal history and structure formation .

- Andrew Liddle *An Introduction to Modern Cosmology*

Another gentle introduction, and one that is especially good when describing expanding spacetimes. However, it becomes more descriptive, and less quantitative, as the subject progresses.

- Scott Dodelson *Modern Cosmology*

Clear, detailed and comprehensive, this is the go-to book for many practising cosmologists.

- Kolb and Turner *The Early Universe*

This book was written before the discovery of anisotropies in the CMB, rendering it rather outdated in some areas. For this reason it should have been removed from reading lists a decade ago. However, the explanations are so lucid, especially when it comes to matters of thermal equilibrium, that few are willing to give it up.

- Steven Weinberg *Cosmology*

Weinberg is one of the smarter Nobel prize winners in physics. Here he offers a scholarly account of the subject, devoid of pretty pictures and diagrams, and with a dogged refusal to draw graphs, yet full of clarity and insight.

A number of further lecture notes are available on the web. Links can be found on the course webpage: http://www.damtp.cam.ac.uk/user/tong/cosmo.html

Contents

0. Introduction 1

1. The Expanding Universe 3
- 1.1 The Geometry of Spacetime 7
  - 1.1.1 Homogeneous and Isotropic Spaces 7
  - 1.1.2 The FRW Metric 10
  - 1.1.3 Redshift 13
  - 1.1.4 The Big Bang and Cosmological Horizons 15
  - 1.1.5 Measuring Distance 19
- 1.2 The Dynamics of Spacetime 23
  - 1.2.1 Perfect Fluids 24
  - 1.2.2 The Continuity Equation 27
  - 1.2.3 The Friedmann Equation 29
- 1.3 Cosmological Solutions 32
  - 1.3.1 Simple Solutions 32
  - 1.3.2 Curvature and the Fate of the Universe 37
  - 1.3.3 The Cosmological Constant 40
  - 1.3.4 How We Found Our Place in the Universe 44
- 1.4 Our Universe 47
  - 1.4.1 The Energy Budget Today 48
  - 1.4.2 Dark Energy 53
  - 1.4.3 Dark Matter 57
- 1.5 Inflation 64
  - 1.5.1 The Flatness and Horizon Problems 64
  - 1.5.2 A Solution: An Accelerating Phase 67
  - 1.5.3 The Inflaton Field 70
  - 1.5.4 Further Topics 75

2. The Hot Universe 77
- 2.1 Some Statistical Mechanics 78
  - 2.1.1 The Boltzmann Distribution 79
  - 2.1.2 The Ideal Gas 81
- 2.2 The Cosmic Microwave Background 84
  - 2.2.1 Blackbody Radiation 84

2.2.2 The CMB Today 88
2.2.3 The Discovery of the CMB 91
2.3 Recombination 93
2.3.1 The Chemical Potential 94
2.3.2 Non-Relativistic Gases Revisited 95
2.3.3 The Saha Equation 98
2.3.4 Freeze Out and Last Scattering 102
2.4 Bosons and Fermions 105
2.4.1 Bose-Einstein and Fermi-Dirac Distributions 106
2.4.2 Ultra-Relativistic Gases 109
2.5 The Hot Big Bang 111
2.5.1 Temperature vs Time 111
2.5.2 The Thermal History of our Universe 116
2.5.3 Nucleosynthesis 117
2.5.4 Further Topics 123

3. Structure Formation 127
3.1 Density Perturbations 128
3.1.1 Sound Waves 128
3.1.2 Jeans Instability 132
3.1.3 Density Perturbations in an Expanding Space 134
3.1.4 The Growth of Perturbations 139
3.1.5 Validity of the Newtonian Approximation 144
3.1.6 The Transfer Function 145
3.2 The Power Spectrum 147
3.2.1 Adiabatic, Gaussian Perturbations 149
3.2.2 The Power Spectrum Today 152
3.2.3 Baryonic Acoustic Oscillations 154
3.2.4 Window Functions and Mass Distribution 155
3.3 Nonlinear Perturbations 159
3.3.1 Spherical Collapse 160
3.3.2 Virialisation and Dark Matter Halos 163
3.3.3 Why the Universe Wouldn't be Home Without Dark Matter 164
3.3.4 The Cosmological Constant Revisited 166
3.4 The Cosmic Microwave Background 167
3.4.1 Gravitational Red-Shift 168
3.4.2 The CMB Power Spectrum 169
3.4.3 A Very Brief Introduction to CMB Physics 171

3.5 Inflation Revisited 173
3.5.1 Superhorizon Perturbations 174
3.5.2 Classical Inflationary Perturbations 175
3.5.3 The Quantum Harmonic Oscillator 178
3.5.4 Quantum Inflationary Perturbations 182
3.5.5 Things We Haven't (Yet?) Seen 185

Acknowledgements

This is an introductory course on cosmology aimed at mathematics undergraduates at the University of Cambridge. You will need to be comfortable with the basics of Special Relativity, but no prior knowledge of either General Relativity or Statistical Mechanics is assumed. In particular, the minimal amount of statistical mechanics will be developed in order to understand what we need. I have made the slightly unusual choice of avoiding all mention of entropy, on the grounds that nearly all processes in the early universe are adiabatic and we can, for the most part, get by without it. (The two exceptions are a factor of 4/11 in the cosmic neutrino background and, relatedly, the number of effective relativistic species during nucleosynthesis: for each of these I've quoted, but not derived, the relevant result about entropy.)

I'm very grateful to both Enrico Pajer and Blake Sherwin for explaining many subtle (and less subtle) points to me, and to Daniel Baumann and Alex Considine Tong for encouragement. I'm supported by the Royal Society and by the Simons Foundation.

0. Introduction

All civilisations have an origin myth. We are the first to have got it right.

Our origin myth goes by the name of the Big Bang theory. It is a wonderfully evocative name, but one that seeds confusion from the off. The Big Bang theory does not say that the universe started with a bang. In fact, the Big Bang theory has nothing at all to say about the birth of the universe. There is a very simple answer to the question "how did the universe begin?" which is "we don't know".

Instead our origin myth is more modest in scope. It tells us only what the universe was like when it was very much younger. Our story starts from a simple observation: the universe is expanding. This means, of course, that in earlier times everything was closer together. We take this observation and push it to the extreme. As objects are forced closer together, they get hotter. The Big Bang theory postulates that there was a time, in the distant past, when the Universe was so hot that matter, atoms and even nuclei melted and all of space was filled with a fireball. The Big Bang theory is a collection of ideas, calculations and predictions that explain what happened in this fireball, and how it subsequently evolved into the universe we see around us today.

The word "theory" in the Big Bang theory might suggest an element of doubt. This is misleading. The Big Bang theory is a "theory" in the same way that evolution is a "theory". In other words, it happened. We know that the universe was filled with a fireball for a very simple reason: we've seen it. In fact, not only have we seen it. We have taken a photograph of it. This photograph, shown in Figure 1, contains a wealth of information about what the universe was like when it was much younger. Of course, this being science we don't like to brag about these things, so rather than jumping up and down and shouting "we've taken a fucking photograph of the fucking Big Bang", we instead wrap it up in dull technical words. We call it the cosmic microwave background radiation. We may, as a community, have underplayed our hand a little here.

As we inch further back towards the "$t = 0$" moment, known colloquially but inaccurately as "the Big Bang", the universe gets hotter and energies involved get higher. One of the goals of cosmology is to push back in time as far as possible to get closer to that mysterious "$t = 0$" moment. Progress here has been nothing short of astonishing. As we will learn, we have a very good idea of what was happening a minute or so after the Big Bang, with detailed calculations of the way different elements are forged in the early universe in perfect agreeement with observations. As we go back further, the observational evidence is harder to come by, but our theories of particle physics give us a reasonable level of confidence back to $t = 10^{-12}$ seconds after the Big Bang. As

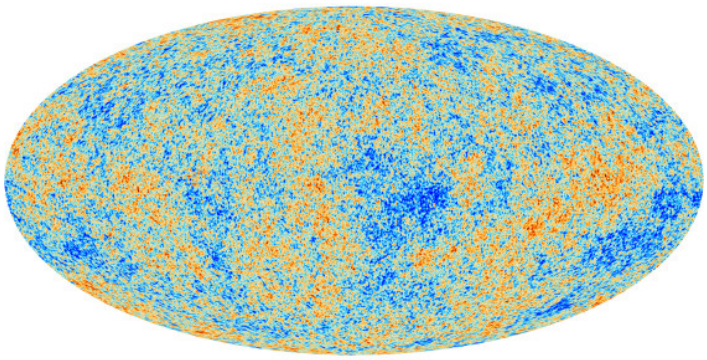

Figure 1: This is a photograph of the Big Bang.

we will see, there are also good reasons to think that, at still earlier times, there was a period of very rapid expansion in the universe known as inflation.

It feels strange to talk with any level of seriousness about the universe when it was a few minutes old, let alone at time $t < 10^{-12}$ seconds. Nonetheless, there are a number of clues surviving in the universe to tell us about these early times, all of which can be explained with impressive accuracy by applying some simple and well tested physical ideas to this most extreme of environments.

The purpose of these lectures is to tell the story above in some detail, to describe 13.8 billion years of history, starting when the Universe was just a fraction of a second old, and extending to the present day.

1. The Expanding Universe

Our goal in this section is an ambitious one: we wish to construct, and then solve, the equations that govern the evolution of the entire universe.

When describing any system in physics, the trick is to focus on the right degrees of freedom. A good choice of variables captures the essence of the problem, while ignoring any irrelevant details. The universe is no different. To motivate our choice, we make the following assumption: the universe is a dull and featureless place. To inject some gravity into this proposal, we elevate it to an important sounding principle:

> The Cosmological Principle: On the largest scales, the universe is spatially homogeneous and isotropic.

Here, *homogeneity* is the property that the universe looks identical at every point in space, while *isotropy* is the property that it looks the same in every direction. Note that the cosmological principle refers only to space. The universe is neither homogenous nor isotropic in time, a fact which underpins this entire course.

Why make this assumption? The primary reason is one of expediency: the universe is, in reality, a complicated place with interesting things happening in it. But these things are discussed in other courses and we will be best served by ignoring them. By averaging over such trifling details, we are left with a description of the universe on the very largest scales, where things are simple.

This averaging ignores little things, like my daily routine, and it is hard to imagine that these have much cosmological significance. However, it also ignores bigger things, like the distribution of galaxies in the universe, that one might think are relevant. Our plan is to proceed with the assumption of simplicity and later, in Section 3, see how we can start to add in some of the details.

The cosmological principle sounds eminently reasonable. Since Copernicus we have known that, while we live in a very special place, we are not at the centre of everything. The cosmological principle allows us to retain our sense of importance by invoking the argument: "if we're not at the centre, then surely no one else is either". You should, however, be suspicious of any grand-sounding principle. Physics is an empirical science and in recent decades we have developed technologies to the point where the cosmological principle can be tested. Fortunately, it stands up pretty well. There are two main pieces of evidence:

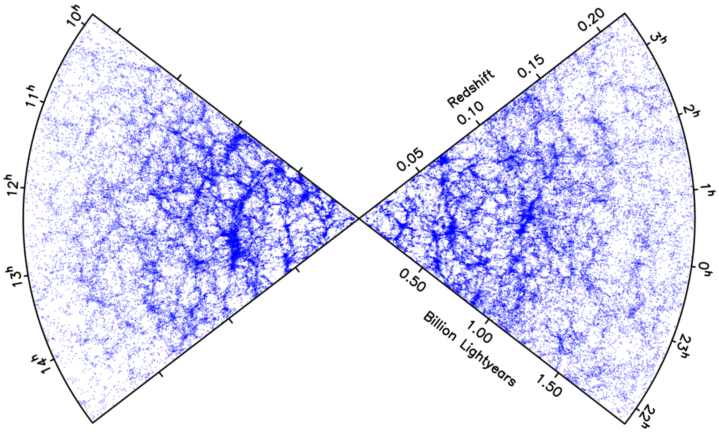

Figure 2: The distribution of galaxies in a wedge in the sky, as measured by the 2dF redshift survey. The distribution looks increasingly smooth on larger scales.

- The cosmic microwave background radiation (CMB) is the afterglow of the Big Bang, an almost uniform sea of photons which fills all of space and provides a snapshot of the universe from almost 14 billion years ago. This is important and will be discussed in more detail in Section 2.2. The temperature of the CMB is[1]

$$T_{\rm CMB} \approx 2.73 \text{ K}$$

  However, it's not quite uniform. There are small fluctuations in temperature with a characteristic scale

$$\frac{\delta T}{T_{\rm CMB}} \sim 10^{-5}$$

  These fluctuations are depicted in the famous photograph shown in Figure 1, taken by the Planck satelite. The fact that the temperature fluctuations are so small is telling us that the early universe was extremely smooth.

- A number of *redshift surveys* have provided a 3d map of hundreds of thousands of galaxies, stretching out to distances of around 2×10^9 light years. The evidence suggests that, while clumpy on small scales, the distribution of galaxies is roughly homogeneous on distances greater than $\sim 3 \times 10^8$ lightyears. An example of such a galaxy survey is shown in Figure 2.

[1]The most accurate determination gives $T_{\rm CMB} = 2.72548 \pm 0.00057$ K; See D.J. Fixsen, "*The Temperature of the Cosmic Microwave Background*", arXiv:0911.1955.

A Sense of Scale

Before we proceed, this is a good time to pause and try to gain some sense of perspective about the universe. First, let's introduce some units. The standard SI units are hopelessly inappropriate for use in cosmology. The metre, for example, is officially defined to be roughly the size of things in my house. Thinking slightly bigger, the average distance from the Earth to the Sun, also known as one *Astronomical Unit* (symbol AU), is

$$1 \text{ AU} \approx 1.5 \times 10^{11} \text{ m}$$

To measure distances of objects that lie beyond our solar system, it's useful to introduce further, farther units. A familiar choice is the lightyear (symbol ly), given by

$$1 \text{ ly} \approx 9.5 \times 10^{15} \text{ m}$$

However, a more commonly used unit among astronomers is the *parsec* (symbol pc), which is based on the observed parallax motion of stars as the Earth orbits the Sun. A parsec is defined as the distance at which a star will exhibit one arcsecond of parallax, which means it wobbles by $1/3600^{\text{th}}$ of a degree in the sky over the course of a year.

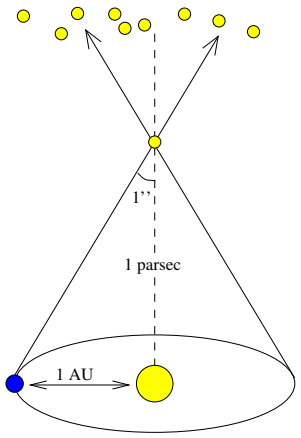

Figure 3: Not to scale.

$$1 \text{ pc} \approx 3.26 \text{ ly}$$

This provides a good unit of measurement to nearby stars. Our closest neighbour, Proxima Centauri, sits at a distance of 1.3 pc. The distance to the centre of our galaxy, the Milky Way, is around 8×10^3 pc, or 8 kpc. Our galaxy is home to around 100 billion stars (give or take) and is approximately 30 kpc across.

The are a large number of neighbouring dwarf galaxies, some of which are actually closer to us than the centre of the Milky Way. But the nearest spiral galaxy is Andromeda, which is approximatey 1 *Megaparsec* (symbol Mpc) or one million parsecs away. The megaparsec is one of the units of choice for cosmologists.

Galaxies are not the largest objects in the universe. They, in turn, gather into clusters and then superclusters and various other filamentary structures. There also appear to be enormous voids in the universe, and it seems plausible that there are more big things to find. Currently, the largest such structures appear to be a few 100 Mpc or so across.

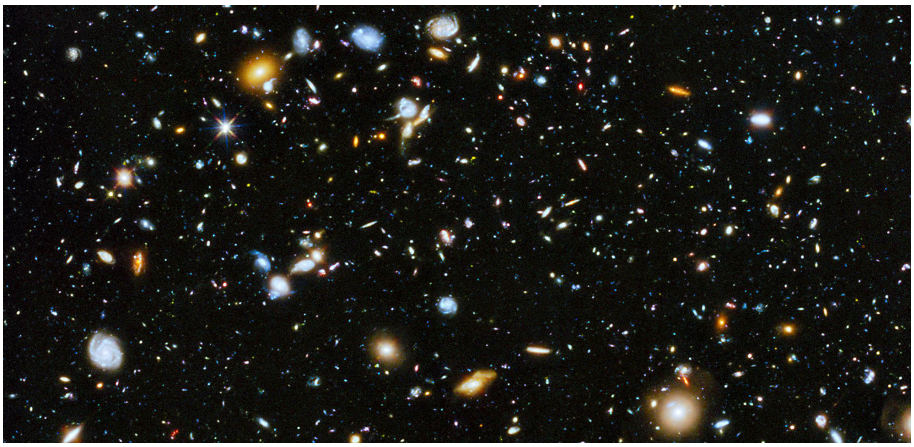

Figure 4: Hubble ultra-deep field shows around 10,000 galaxies.

All of this is to say that we have to look at very large scales before the universe appears to obey the cosmological principle, but it does finally get there. As we will see later in this course, there is a limit to how big we can go. The size of the observable universe is around

$$3000 \text{ Mpc} \approx 10^{26} \text{ m}$$

and there seems to be no way to peer beyond this. The observable universe contains, we think, around 100 billion galaxies, each of them with around 100 billion stars. It is difficult to build intuition for numbers this big, and distances this vast. Some help comes from the Hubble ultra-deep field, shown in Figure 4, which covers a couple of arcminutes of sky, roughly the same as a the tip of a pencil held out at arms length. The image shows around 10,000 galaxies, some no more than a single pixel, but each containing around 100 billion suns, each of which is likely to play host to a solar system of planets.

For more intuition about the size of the universe, we turn to the classics

> "When you're thinking big, think bigger than the biggest thing ever and then some. Much bigger than that in fact, really amazingly immense, a totally stunning size, real 'wow, that's big' time. It's just so big that by comparison, bigness itself looks really titchy. Gigantic multiplied by colossal multiplied by staggeringly huge is the sort of concept we're trying to get across here."
>
> *Douglas Adams*

1.1 The Geometry of Spacetime

The cosmological principle motivates us to treat the universe as a boring, featureless object. Given this, it's not obvious what property of the universe we have left to focus on. The answer is to be found in geometry.

1.1.1 Homogeneous and Isotropic Spaces

The fact that space (and time) can deviate from the seemingly flat geometry of our everyday experience is the essence of the theory general relativity. Fortunately, we will need very little of the full theory for this course. This is, in large part, due to the cosmological principle which allows us to focus on spatial geometries which are homogeneous and isotropic. There are three such geometries:

- **Flat Space:** The simplest homogeneous and isotropic three-dimensional space is flat space, also known as Euclidean space. We will denote it by $\mathbf{R}^3$.

  We describe the geometry of any space in terms of a *metric*. This gives us a prescription for measuring the distance between two points on the space. More precisely, we will specify the metric in terms of the *line element* ds which tells us the infinitesimal distance between two nearby points. For flat space, this is the familiar Euclidean metric

  $$ds^2 = dx^2 + dy^2 + dz^2 \tag{1.1}$$

  We'll also work in a number of other coordinates systems, such as spherical polar coordinates

  $$x = r\sin\theta\cos\phi \quad , \quad y = r\sin\theta\sin\phi \quad , \quad z = r\cos\theta \tag{1.2}$$

  with $r \in [0, \infty)$, $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi)$. To compute the metric in these coordinates, we relate small changes in (r, θ, ϕ) to small changes in (x, y, z) by the Leibniz rule, giving

  $$\begin{aligned} dx &= dr\sin\theta\cos\phi + r\cos\theta\cos\phi\, d\theta - r\sin\theta\sin\phi\, d\phi \\ dy &= dr\sin\theta\sin\phi + r\cos\theta\sin\phi\, d\theta + r\sin\theta\cos\phi\, d\phi \\ dz &= dr\cos\theta - r\sin\theta\, d\theta \end{aligned}$$

  Substituting these expressions into the flat metric (1.1) gives us the flat metric in polar coordinates

  $$ds^2 = dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \tag{1.3}$$

- **Positive Curvature** The next homogeneous and isotropic space is also fairly intuitive: we can take a three-dimensional sphere $\mathbf{S}^3$, constructed as an embedding in four-dimensional Euclidean space $\mathbf{R}^4$

$$x^2 + y^2 + z^2 + w^2 = R^2$$

with R the radius of the sphere. The sphere has uniform positive curvature. On such a space, parallel lines will eventually meet.

We again have different choices of coordinates. One option is to retain the 3d spherical polars (1.2) and eliminate w using $w^2 = R^2 - r^2$. A point on the sphere $\mathbf{S}^3$ is then labelled by a "radial" coordinate r, with range $r \in [0, R]$, and the two angular coordinates $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi)$. We can compute the metric on $\mathbf{S}^3$ by noting

$$w^2 = R^2 - r^2 \quad \Rightarrow \quad dw = -\frac{r\, dr}{\sqrt{R^2 - r^2}}$$

The metric on the sphere is then inherited from the flat metric in $\mathbf{R}^4$. We substitute the expression above into the flat metric $ds^2 = dx^2 + dy^2 + dz^2 + dw^2$ to find the metric on $\mathbf{S}^3$,

$$ds^2 = \frac{R^2}{R^2 - r^2} dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \tag{1.4}$$

Strictly speaking, this set of coordinates only covers half the $\mathbf{S}^3$, the hemisphere with $w \geq 0$.

Arguably a more natural set of coordinates are provided by the 4d generalisation of the spherical polar coordinates (1.2). These are defined by writing $r = R\sin\chi$, so

$$\begin{aligned} x = R\sin\chi\sin\theta\cos\phi \quad &, \quad y = R\sin\chi\sin\theta\sin\phi \\ z = R\sin\chi\cos\theta \quad &, \quad w = R\cos\chi \end{aligned} \tag{1.5}$$

Now a point on $\mathbf{S}^3$ is determined by three angular coordinates, $\chi, \theta \in [0, \pi]$ and $\phi \in [0, 2\pi)$. The metric becomes

$$ds^2 = R^2\Big[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta\, d\phi^2)\Big] \tag{1.6}$$

Although we introduced the 3d sphere $\mathbf{S}^3$ by embedding it $\mathbf{R}^4$, the higher dimensional space is a crutch that we no longer need. Worse, it is a crutch that can

be quite misleading. Both mathematically, and physically, the sphere $\mathbf{S}^3$ makes sense on its own without any reference to a space in which it's embedded. In particular, should we discover that the spatial geometry of our universe is $\mathbf{S}^3$, this does not imply the physical existence of some ethereal $\mathbf{R}^4$ in which the universe is floating.

- **Negative Curvature** Our final homogeneous and isotropic space is perhaps the least familiar. It is a hyperboloid $\mathbf{H}^3$, which can again be defined as an embedding in $\mathbf{R}^4$, this time with

$$x^2 + y^2 + z^2 - w^2 = -R^2 \tag{1.7}$$

This is a space of uniform negative curvature. Parallel lines diverge on a space with negative curvature.

Once again, the metric is inherited from the embedding in $\mathbf{R}^4$, but this time with signature $(+++-)$, so $ds^2 = dx^2 + dy^2 + dz^2 - dw^2$ as befits the embedding (1.7). Using the 3d coordinates (r, θ, ϕ), we have $w^2 = r^2 + R^2$. The metric is

$$ds^2 = \frac{R^2}{R^2 + r^2} dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \tag{1.8}$$

Alternatively, we can write $r = R\sinh\chi$, in which case the metric becomes

$$ds^2 = R^2\Big[d\chi^2 + \sinh^2\chi(d\theta^2 + \sin^2\theta\, d\phi^2)\Big] \tag{1.9}$$

It is often useful to write these metrics in a unified form. In the (r, θ, ϕ) coordinates, we can write the general metric (1.3), (1.4) and (1.8) as

$$ds^2 = \frac{dr^2}{1 - kr^2/R^2} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \quad \text{with } k = \begin{cases} +1 & \text{Spherical} \\ 0 & \text{Euclidean} \\ -1 & \text{Hyperbolic} \end{cases} \tag{1.10}$$

Throughout these lectures, we will use $k = -1, 0, +1$ to denote the three possible spatial geometries. Alternatively, in the coordinates (χ, θ, ϕ), the metrics (1.3), (1.6) and (1.9) can be written in a unified way as

$$ds^2 = R^2\Big[d\chi^2 + S_k^2(\chi)(d\theta^2 + \sin^2\theta\, d\phi^2)\Big] \quad \text{with } S_k(\chi) = \begin{cases} \sin\chi & k = +1 \\ \chi & k = 0 \\ \sinh\chi & k = -1 \end{cases} \tag{1.11}$$

where now χ is a dimensionless coordinate. (In flat space, we have to introduce an arbitrary, fiducial scale R to write the metric in this form.)

Global Topology

We have identified three possible spatial geometries consistent with the cosmological principle. Of these, $\mathbf{S}^3$ is a *compact* space, meaning that it has finite volume (which is $2\pi^2 R^3$). In contrast, both $\mathbf{R}^3$ and $\mathbf{H}^3$ are non-compact, with infinite volume.

In fact, it is straightforward to construct compact spaces for the $k = 0$ and $k = -1$ cases. We simply need to impose periodicity conditions on the coordinates. For example, in the $k = 0$ case we could identify the points $x^i = x^i + R^i$, $i =, 1, 2, 3$ with some fixed R^i. This results in the torus $\mathbf{T}^3$.

Spaces constructed this way are homogenous, but no longer isotropic. For example, on the torus there are special directions that bring you back to where you started on the shortest path. This means that such spaces violate the cosmological principle. More importantly, there is no observational evidence that they do, in fact, describe our universe so we will not discuss them in what follows.

1.1.2 The FRW Metric

Our universe is not three-dimensional. It is four-dimensional, with time as the forth coordinate. In special relativity, we consider the flat four-dimensional spacetime known as Minkowski space, with metric[2]

$$ds^2 = -c^2\, dt^2 + d\mathbf{x}^2$$

with c the speed of light. This metric has the property that the distance between two points in spacetime is invariant under Lorentz transformations; it is the same for all inertial observers.

The Minkowski metric is appropriate for describing physics in some small region of space and time, like the experiments performed here on Earth. But, on cosmological scales, the Minkowski metric needs replacing so that it captures the fact that the universe is expanding. This is straightforward. We replace the flat spatial metric $d\mathbf{x}^2$ with one of the three homogeneous and isotropic metrics that we met in the previous section and write

$$ds^2 = -c^2\, dt^2 + a^2(t)\left[\frac{1}{1 - kr^2/R^2}dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)\right] \tag{1.12}$$

This is the Friedmann-Robertson-Walker, or FRW metric. The role of the dimensionless *scale factor* $a(t)$ is, as we shall see, to change distances over time.

[2]An introduction to special relativity can be found in Section 7 of the lectures on Dynamics and Relativity. There we used the metric with opposite signature $ds^2 = +c^2\, dt^2 - d\mathbf{x}^2$.

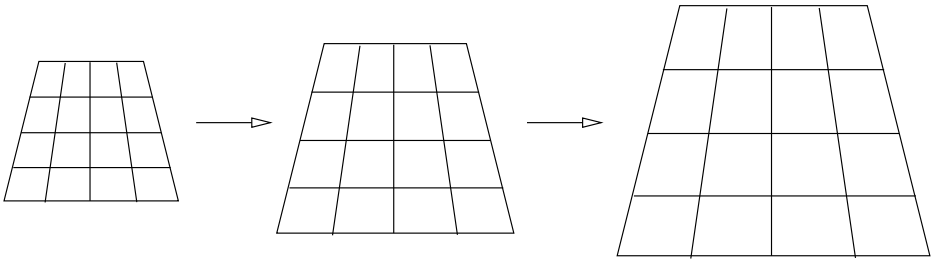

Figure 5: The expansion of the universe. The physical distance between fixed co-moving coordinates increases with time.

There is a redundancy in the description of the metric. If we rescale coordinates as $a \to \lambda a$, $r \to r/\lambda$ and $R \to R/\lambda$ then the metric remains unchanged. We use this to set the scale factor evaluated at the present time t_0 to unity,

$$a_0 = a(t_0) = 1$$

where the subscript 0 will always denote the value of a quantity evaluated today.

Consider a galaxy sitting at some fixed point point (r, θ, ϕ). We refer to the coordinates (r, θ, ϕ) (or, equivalently, (χ, θ, ϕ)) on the 3d space as *co-moving coordinates*. They are analogous to the Lagrangian coordinates used in fluid mechanics. The physical (or proper) distance between the point (r, θ, ϕ) and the origin is then

$$d_{\rm phys} = a(t) \int_0^r \frac{1}{\sqrt{1 - kr'^2/R^2}}\, dr' = a(t) R\chi \tag{1.13}$$

However, there is nothing special about the origin, and the same scaling with $a(t)$ is seen for the distance between any two points. If we choose a function $a(t)$ with $\dot{a} > 0$, then the distance between any two points is increasing. This is the statement that the universe is expanding: two galaxies, at fixed co-moving co-ordinates, will be swept apart as spacetime stretches.

Importantly, the universe isn't expanding "into" anything. Instead, the geometry of spacetime, as described by the metric (1.12), is getting bigger, without reference to anything which sits outside. Similarly, a metric with $\dot{a} < 0$ describes a contracting universe. In Section 1.2, we will introduce the tools needed to calculate $a(t)$. But first, we look at some general features of expanding, or contracting universes.

The FRW metric is not invariant under Lorentz transformation. This means that the universe picks out a preferred rest frame, described by co-moving coordinates. We

can still shift this rest frame by translations (in flat space) or rotations, but not by Lorentz boosts. Consider a galaxy which, in co-moving coordinates, traces a trajectory $\mathbf{x}(t)$. Then, in physical coordinates, the position is

$$\mathbf{x}_{\rm phys}(t) = a(t)\mathbf{x}(t) \tag{1.14}$$

The physical velocity is then

$$\mathbf{v}_{\rm phys}(t) = \frac{d\mathbf{x}_{\rm phys}}{dt} = \frac{da}{dt}\mathbf{x} + a\frac{d\mathbf{x}}{dt} = H\mathbf{x}_{\rm phys} + \mathbf{v}_{\rm pec} \tag{1.15}$$

There are two terms. The first, which is due entirely to the expansion of the universe is written in terms of the *Hubble parameter*,

$$H(t) = \frac{\dot{a}}{a}$$

The second term, $\mathbf{v}_{\rm pec}$, is referred to as the *peculiar velocity* and is describes the inherent motion of the galaxy relative to the cosmological frame, typically due to the gravitational attraction of other nearby galaxies.

Our own peculiar velocity is $v_{\rm pec} \approx 400$ km s^{-1} which is pretty much typical for a galaxy. Meanwhile, the present day value of the Hubble parameter is

$$H_0 \approx 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$$

This is, rather misleadingly, referred to as the *Hubble constant*. Clearly there is nothing constant about it. Although, in fairness, it is pretty much the same today as it was yesterday. It is also common to see the notation

$$H_0 = 100h \text{ km s}^{-1} \text{ Mpc}^{-1} \tag{1.16}$$

and then to describe the value of the Hubble constant in terms of the dimensionless number $h \approx 0.7$. In this course, we'll simply use the notation H_0.

The Hubble parameter has dimensions of time^{-1}, but is written in the rather unusual units km s^{-1} Mpc^{-1}. This is telling us that a galaxy 1 Mpc away will be seen to be retreating at a speed of 70 km s^{-1} due to the expansion of space. For nearby galaxies, this tends to be smaller than their peculiar velocity. However, as we look further away, the expansion term will dominate. The numbers above suggest that this will happen at distances around $400/70 \approx 5$ Mpc.

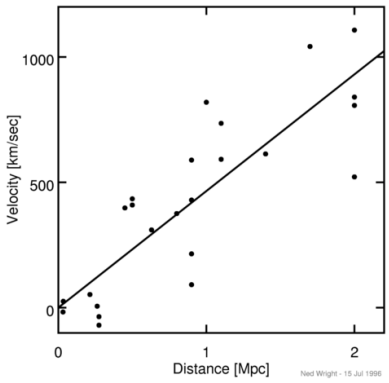

Figure 6: Hubble's original data, from 1929, with a rather optimistic straight line drawn through it.

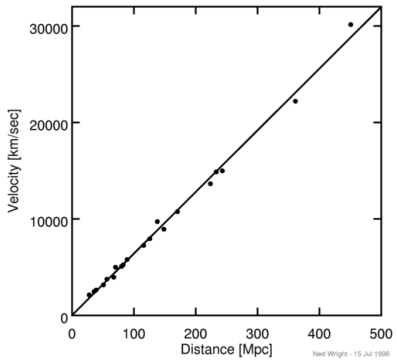

Figure 7: Data from 1996, looking out to much further distances.

If we ignore the peculiar velocities, and further assume that we can approximate the Hubble parameter $H(t)$ as the constant H_0, then the velocity law (1.15) becomes a linear relation between velocity and distance

$$\mathbf{v}_{\text{phys}} = H_0\, \mathbf{x}_{\text{phys}} \tag{1.17}$$

This linear relationship is referred to as Hubble's law; some data is shown in the figures[3]. At yet further distances, we would expect the time dependence of $H(t)$ to reveal itself. We will discuss this in Section 1.4.

There is no obstacle in (1.17) to velocities that exceed the speed of light, $|\mathbf{v}_{\text{phys}}| > c$. This may make you nervous. However, there is no contradiction with relativity and, indeed, the entire framework that we have discussed above sits, without change, in the full theory of general relativity. The statement that "nothing can travel faster than the speed of light" is better thought of as "nothing beats light in a race". Given two objects at the same point, their relative velocity is always less than c. However, the velocity $\mathbf{v}_{\text{phys}}$ is measuring the relative velocity of two objects at very distant points and, in an expanding spacetime, there is no such restriction.

1.1.3 Redshift

All our observational information about the universe comes to us through light waves and, more recently, gravitational waves. To correctly interpret what we're seeing, we need to understand how such waves travel in an expanding spacetime.

[3]Both of these plots are taken from Ned Wright's cosmology tutorial.

In a spacetime metric, light travels along null paths with $ds = 0$. In the FRW metric (1.12), light travelling in the radial direction (i.e. with fixed θ and ϕ) will follow a path,

$$c\,dt = \pm a(t)\frac{dr}{\sqrt{1-kr^2/R^2}} \tag{1.18}$$

If we place ourselves at the origin, the minus sign describes light moving towards us. Aliens on a distant planet, tuning in for the latest Buster Keaton movie, should use the plus sign.

Suppose that a distant galaxy sits stationary in co-moving coordinate r_1 and emits light at time t_1. We observe this signal at $r = 0$, at time t_0, determined by solving the integral equation

$$c\int_{t_1}^{t_0}\frac{dt}{a(t)} = \int_0^{r_1}\frac{dr}{\sqrt{1-kr^2/R^2}}$$

If the galaxy emits a second signal at time $t_1 + \delta t_1$, this is observed at $t_0 + \delta t_0$, with

$$c\int_{t_1+\delta t_1}^{t_0+\delta t_0}\frac{dt}{a(t)} = \int_0^{r_1}\frac{dr}{\sqrt{1-kr^2/R^2}}$$

The right-hand side of both of these equations is the same because it is written in co-moving coordinates. We therefore have

$$\int_{t_1+\delta t_1}^{t_0+\delta t_0}\frac{dt}{a(t)} - \int_{t_1}^{t_0}\frac{dt}{a(t)} = 0 \quad \Rightarrow \quad \frac{\delta t_1}{a(t_1)} = \frac{\delta t_0}{a(t_0)} = \delta t_0 \tag{1.19}$$

where, in the last equality, we've used the fact that we observe the signal today, where $a(t_0) = 1$. We see that the expansion of the universe means that the time difference between the two emitted signals differs from the time difference between the two observed signals. This has an important implication when applied to the wave nature of light. Two successive wave crests are separated by a time

$$\delta t_1 = \frac{\lambda_1}{c}$$

with λ_1 the wavelength of the emitted light. Similarly, the time interval between two observed wave crests is

$$\delta t_0 = \frac{\lambda_0}{c}$$

The result (1.19) tells us that the wavelength of the observed light differs from that of the emitted light,

$$\lambda_0 = \frac{a(t_0)}{a(t_1)}\lambda_1 = \frac{\lambda_1}{a(t_1)} \tag{1.20}$$

This is intuitive: the light is stretched by the expansion of space as it travels through it so that the observed wavelength is longer than the emitted wavelength. This effect is known as *cosmological redshift*. It shares some similarity with the Doppler effect, in which the wavelength of light or sound from moving sources is shifted. However, the analogy is not precise: the Doppler effect depends only on the relative velocity of the source and emitter, while the cosmological redshift is independent of $\dot{a}$, instead depending on the overall expansion of space over the light's journey time.

The *redshift parameter* z is defined as the fractional increase in the observed wavelength,

$$z = \frac{\lambda_0 - \lambda_1}{\lambda_1} = \frac{1 - a(t_1)}{a(t_1)} \quad \Rightarrow \quad 1 + z = \frac{1}{a(t_1)} \tag{1.21}$$

As this course progresses, we will often refer to times in the past in terms of the redshift z. Today we sit at $z = 0$. When $z = 1$, the universe was half the current size. When $z = 2$, the universe was one third the current size.

The redshift is something that we can directly measure. Light from far galaxies come with a fingerprint, the spectral absorption lines that reveal the molecular and atomic makeup of the stars within. By comparing the frequencies of those lines to those on Earth, it is a simple matter to extract z. As an aside, by comparing the relative positions of spectral lines, one can also confirm that atomic physics in far flung places works the same as on Earth, with no detected changes in the laws of physics or the fundamental constants of nature.

1.1.4 The Big Bang and Cosmological Horizons

We will find that all our cosmological models predict a time in the past, $t_{BB} < t_0$, where the scale factor vanishes, $a(t_{BB}) = 0$. This point is colloquially referred to as the Big Bang. The Big Bang is not a point in space, but is a point in time. It happens everywhere in space.

We can get an estimate for the age of the Universe by Taylor expanding $a(t)$ about the present day, and truncating at linear order. Recalling that $a(t_0) = 1$, we have

$$a(t) \approx 1 + H_0(t - t_0) \tag{1.22}$$

This rather naive expansion suggests that the Big Bang occurs at

$$t_0 - t_{BB} = H_0^{-1} \approx 4.4 \times 10^{17} \text{ s} \approx 1.4 \times 10^{10} \text{ years} \tag{1.23}$$

This result of 14 billion years is surprisingly close to the currently accepted value of around 13.8 billion years. However, there is a large dose of luck in this agreement, since the linear approximation (1.22) is not very good when extrapolated over the full age of the universe. We'll revisit this in Section 1.4.

Strictly speaking, we should not trust our equations at the point $a(t_{BB}) = 0$. The metric (1.12) is singular here, and any matter in the universe will be squeezed to infinite density. In such a regime, our simple minded classical equations are not to be trusted, and should be replaced by a quantum theory of matter and gravity. Despite much work, it remains an open problem to understand the origin of the universe at $a(t_{BB}) = 0$. Did time begin here? Was there a previous phase of a contracting universe? Did the universe emerge from some earlier, non-geometric form? We simply don't know.

Understanding the Big Bang is one of the ultimate goals of cosmology. In the meantime, the game is to push as far back in time as we can, using the classical (and semi-classical) theory of gravity that we trust. We will be able to reach scales $a \ll 1$, even if we can't get all the way to $a = 0$, and follow the subsequent evolution of the universe from the initial hot, dense state to the world we see today. This set of ideas, is often referred to as the *Big Bang theory*, even though it tells us nothing about the initial "Big Bang" itself.

The Size of the Observable Universe

The existence of a special time, t_{BB}, means that there is a limit as to how far we can peer into the past. In co-moving coordinates, the greatest distance $r_{\rm max}$ that we can see is the distance that light has travelled since the Big Bang. From (1.18), this is given by

$$c\int_{t_{BB}}^{t} \frac{dt'}{a(t')} = \int_0^{r_{\rm max}(t)} \frac{dr}{\sqrt{1-kr^2/R^2}}$$

The corresponding physical distance is

$$d_H(t) = a(t)\int_0^{r_{\rm max}(t)} \frac{dr}{\sqrt{1-kr^2/R^2}} = c\,a(t)\int_0^t \frac{dt'}{a(t')} \tag{1.24}$$

This is the size of the observable universe. Note that this size is not simply $c(t - t_{BB})$, which is the naive distance that light has travelled since the Big Bang. Indeed, mathematically it could be that the integral on the left-hand side of (1.24) does not converge at t_{BB}, in which case the maximum distance $r_{\rm max}$ would be infinite.

The distance d_H is sometimes referred to as the *particle horizon*. The name mimics the event horizon of black holes. Nothing inside the event horizon of a black hole can influence the world outside. Similarly, nothing outside the particle horizon can influence us today.

The Event Horizon

> "It does seem rather odd that two or more observers, even such as sat on the same school bench in the remote past, should in future, when they have followed different paths in life, experience different worlds, so that eventually certain parts of the experienced world of one of them should remain by principle inaccessible to the other and vice versa."
>
> Erwin Schrödinger, 1956

The particle horizon tells us that there are parts of the universe that we cannot presently see. One might expect that, as time progresses, more and more of spacetime comes into view. In fact, this need not be the case.

One option is that the universe begins collapsing in the future, and there is a second time $t_{BC} > t_0$ where $a(t_{BC}) = 0$. This is referred to as the Big Crunch. In this case, there is a limit on how far we can communicate before the universe comes to an end, given by

$$c\int_t^{t_{BC}} \frac{dt'}{a(t')} = \int_0^{r_{\rm max}(t)} \frac{dr}{\sqrt{1-kr^2/R^2}}$$

Perhaps more surprisingly, even if the universe continues to expand and the FRW metric holds for $t \to \infty$, then there could still be a maximum distance that we can influence. The relevant equation is now

$$c\int_t^{\infty} \frac{dt'}{a(t')} = \int_0^{r_{\rm max}(t)} \frac{dr}{\sqrt{1-kr^2/R^2}} \tag{1.25}$$

The maximum co-moving distance $r_{\rm max}$ is finite provided that the left-hand side converges. For example, this happens if we have $a(t) \sim e^{Ht}$ as $t \to \infty$. As we will see later in the course, this seems to be the most likely fate of our universe. As Schrödinger described, it is quite possible that two friends who once played together as children could move apart from each other, only to find that they've travelled too far and can never return as they are inexorably swept further apart by the expansion of the universe. It's not a bad metaphor for life.

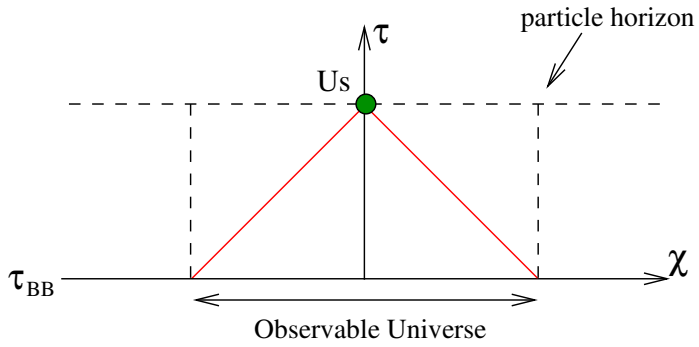

Figure 8: The particle horizon defines the size of your observable universe.

In this context, the distance $r_{\rm max}(t)$ is called the (co-moving) *cosmological event horizon*. Once again, there is the analogy with the black hole. Regions beyond the cosmological horizon are beyond our reach; if we choose to sit still, we will never see them and never communicate with them. However, there are also important distinctions. In contrast to the event horizon of a black hole, the concept of cosmological event horizon depends on the choice of observer.

Conformal Time

The properties of horizons are perhaps best illustrated by introducing a different time coordinate,

$$\tau = \int^t \frac{dt'}{a(t')} \tag{1.26}$$

This is known as *conformal time*. If we also work with the χ spatial coordinate (1.11) then the FRW metric takes the simple form

$$ds^2 = a^2(\tau)\left[-c^2d\tau^2 + R^2d\chi^2 + R^2S_k(\chi)^2(d\theta^2 + \sin^2\,d\phi^2)\right]$$

with all time dependence sitting as on overall factor outside. This has a rather nice consequence because if we draw events in the $(c\tau, R\chi)$ plane then light-rays, which travel with $ds^2 = 0$, correspond to 45° lines, just like in Minkowski space. This helps visualise the causal structure of an expanding universe.

Suppose that we sit at some conformal time τ. A signal can be emitted no earlier than τ_{BB} where the Big Bang singularity occurs. This then puts a restriction on how far we can see in space, defined to be the particle horizon

$$R\chi_{\rm ph} = c(\tau - \tau_{BB})$$

This is shown in Figure 8.

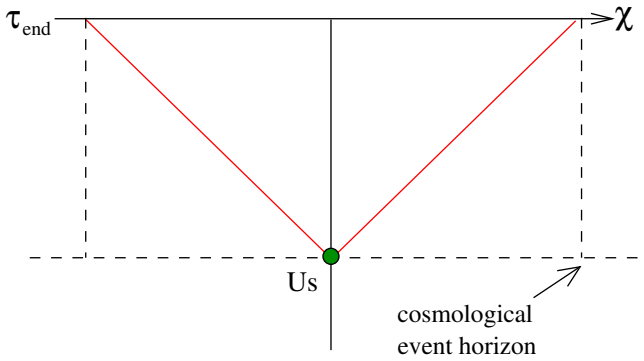

Figure 9: The cosmological event horizon defines the events you can hope to influence.

Looking forward, the issue comes because the end of the universe, at $t \to \infty$, corresponds to a finite conformal time $\tau_{\rm end}$. This means that nothing we can do will be seen beyond a maximum distance which defines the cosmological event horizon,

$$R\chi_{\rm eh} = \tau_{\rm end} - \tau$$

This is shown in Figure 9.

It turns out that conformal time is also a useful change of variable when solving the equations of cosmology. We'll see an example in Section 1.3.2.

1.1.5 Measuring Distance

These lectures are unapologetically theoretical. Nonetheless, we should ask how we know certain facts about the universe. One of the most important challenges facing observational astronomers and cosmologists is the need to accurately determine the distance to various objects in the universe. This is crucial if we are to reconstruct the history of the expansion of the universe $a(t)$.

Furthermore, there is even an ambiguity in what we mean by "distance". So far, we have defined the co-moving distance $R\chi$ and, in (1.13), the physical distance $d_{\rm phys}(t) = a(t)R\chi$. The latter is, as the name suggests, more physical, but it does not equate directly to something we can measure. Instead, $d_{\rm phys}(t)$ is the distance between two events which took place at some fixed time t, but to measure this distance, we would need to pause the expansion of the universe while we wheeled out a tape measure, typically one which stretches over several megaparsecs. This, it turns out, is impractical.

Figure 10: This cow is small.

Figure 11: This cow is far away.

For these reasons, we need a more useful definition of distance and how to measure it. A useful measure of distance should involve what we actually see, and what we see is light that has travelled across the universe, sometimes for a long long time.

For objects that are reasonably close, we can use parallax, the slight wobble of a stars position caused by the Earth orbiting the Sun. The current state of the art is the Gaia satellite which can measure the parallax of sufficiently bright star sto an accuracy of 2×10^{-5} arc seconds, corresponding to distances of $1/10^{\text{th}}$ of a megaparsec. While impressive this is, to quote the classics, peanuts to space. We therefore need to turn to more indirect methods.

The Luminosity Distance

One way to measure distance is to use the brightness of the object. Obviously, the further away an object is, the less bright it appears in the sky. The problem with this approach is that it's difficult to be sure if an object is genuinely far away, or intrinsically dim. It is entirely analogous to the famous problem with cows: how do we tell if they are small, or merely far away?

To resolve this degeneracy, cosmologists turn to *standard candles*. These are objects whose intrinsic brightness can be determined by other means. There are a number of candidates for standard candles, but some of the most important are:

- Cepheids are bright stars which pulsate with a period ranging from a few days to a month. This periodicity is thought to vary linearly with the intrinsic brightness of the star. These were the standard candles originally used by Hubble.

- A type Ia supernova arises when a white dwarf accretes too much matter from an orbiting companion star, pushing it over the Chandrasekhar limit (the point at which a star collapses). Such events are rare — typically a few a century in a galaxy the size of the Milky Way — but with a brightness that is comparable

to all the stars in the host galaxy. The universal nature of the Chandrasekhar limit means that there is considerable uniformity in these supernovae. What little variation there is can be accounted for by studying the "light curve", meaning how fast the supernova dims after the original burst. These supernovae were first developed as standard candles in the 1990s and resulted in the discovery of the acceleration of the universe.

- The more recent discovery of gravitational waves opens up the possibility for a *standard siren*. The gravitational waveform can be used to accurately determine the distance. When these waves arise from the collision of a neutron star and black hole (sometimes called a kilanova), the event can also be seen in the electromagnetic spectrum, allowing identification of the host galaxy.

Given a standard candle, we can be fairly sure that we know the intrinsic *luminosity* L of an object, defined as the energy emitted per unit time. We would like to determine the apparent luminosity l, defined by the energy per unit time per unit area, seen by a distant observer. In flat space, this is straightforward: at a distance d, the energy has spread out over a sphere $\mathbf{S}^2$ of area $4\pi d^2$, giving us

$$l = \frac{L}{4\pi d^2} \quad \text{in flat space} \tag{1.27}$$

The question we would like to ask is: how does this generalise in an FRW universe? To answer this, it's best to work in the coordinates (1.11), so the FRW metric reads

$$ds^2 = -c^2dt^2 + R^2\Big[d\chi^2 + S_k^2(\chi)(d\theta^2 + \sinh^2\theta\, d\phi^2)\Big]$$

with

$$S_k(\chi) = \begin{cases} \sin\chi & k = +1 \\ \chi & k = 0 \\ \sinh\chi & k = -1 \end{cases}$$

There are now three things that we need to take into account. The first is that a sphere $\mathbf{S}^2$ with radius χ now has area $4\pi R^2 S_k(\chi)^2$, which agrees with our previous result in flat space, but differs when $k \neq 0$. Secondly, the photons are redshifted after their long journey. If they are emitted with frequency ν_1 then, from (1.20), they arrive with frequency

$$\nu_0 = \frac{2\pi c}{\lambda_0} = \frac{\nu_1}{1+z}$$

This lower arrival rate decreases the observed flux. Finally, the observed energy E_0 of each photon is reduced compared to the emitted energy E_1,

$$E_0 = \hbar\nu_0 = \frac{E_1}{1+z}$$

The upshot is that, in an expanding universe, the observed flux from a source with intrinsic luminosity L sitting at co-moving distance χ is

$$l = \frac{L}{4\pi R^2 S_k(\chi)^2(1+z)^2}$$

Comparing to (1.27) motivates us to define the *luminosity distance*

$$d_L(\chi) = RS_k(\chi)(1+z) \tag{1.28}$$

For a standard candle, where L is known, the luminosity distance d_L is something that can be measured. From this, and the redshift, we can infer the co-moving distance $RS_k(\chi)$. In flat space, this is simply $R\chi = r$.

Extracting H_0

Finally, we can use this machinery to determine the Hubble constant H_0. We first Taylor expand the scale factor $a(t)$ about the present day. Setting $a_0 = 1$, we have

$$a(t) = 1 + H_0(t-t_0) - \frac{1}{2}q_0 H_0^2(t-t_0)^2 + \ldots \tag{1.29}$$

Here we've introduced the second order term, with dimensionless parameter q_0. This is known as the *deceleration parameter*, and should be thought of as the present day value of the function

$$q(t) = -\frac{\ddot{a}a}{\dot{a}^2} = -\frac{\ddot{a}}{aH^2}$$

The name is rather unfortunate because, as we will learn in Section 1.4, the expansion of our universe is actually accelerating, with $\ddot{a} > 0$! In our universe, the deceleration parameter is negative: $q_0 \approx -0.5$.

First, we integrate the path of a light-ray (1.18) to get an expression for the co-moving distance χ in terms of the "look-back time" $(t_0 - t_1)$

$$\begin{aligned} R\chi = c\int_{t_1}^{t_0}\frac{dt}{a(t)} &= c\int_{t_1}^{t_0}\Big[1 - H_0(t_0-t) + \ldots\Big]dt \\ &= c(t-t_0)\Big[1 + \frac{1}{2}H_0(t-t_0) + \ldots\Big] \end{aligned} \tag{1.30}$$

Next, we get an expression for the look-back time $t_0 - t_1$ in terms of the redshift z. From (1.21), light emitted at some time t_1 suffers a redshift $1 + z = 1/a(t)$. Inverting the Taylor expansion (1.29), we have

$$z = \frac{1}{a(t_1)} - 1 \approx H_0(t_0 - t_1) + \frac{1}{2}(2 + q_0)H_0^2(t_0 - t_1)^2 + \ldots$$

We now invert this to give the "look-back time" $t_0 - t_1$ as a Taylor expansion in the redshift z. (As an aside: you could do the inversion by solving the quadratic formula, and subsequently Taylor expanding the square-root. But when inverting a power series, it's more straightforward to write an ansatz $H_0(t_0 - t_1) = A_1 z + A_2 z^2 + \ldots$, which we substitute this into the right-hand side and match terms.) We find

$$H_0(t_0 - t_1) = z - \frac{1}{2}(2 + q_0)z^2 + \ldots \tag{1.31}$$

Combining (1.30) and (1.31) gives

$$\frac{H_0 R\chi}{c} = z - \frac{1}{2}(1 + q_0)z^2 + \ldots$$

We can now substitute this into our expression for the luminosity distance (1.28). Life is easiest in flat space, where $RS_k(\chi) = R\chi$ and we find

$$d_L = \frac{c}{H_0}\left(z + \frac{1}{2}(1 - q_0)z^2 + \ldots\right)$$

This expression is valid only for $z \ll 1$. By plotting the observed d_L vs z, and fitting to this functional form, we can extract H_0 and q_0.

1.2 The Dynamics of Spacetime

We have learned that, on the largest distance scales, the universe is described by the FRW metric

$$ds^2 = -c^2\, dt^2 + a^2(t)\left[\frac{1}{1 - kr^2/R^2}dr^2 + r^2(d\theta^2 + \sin^2\, d\phi^2)\right]$$

with the history of the expansion (or contraction) of the universe captured by the function $a(t)$. Our goal now is to calculate this function.

A good maxim for general relativity is: spacetime tells matter how to move, matter tells space how to curve. We saw an example of the first statement in the previous section, with galaxies swept apart by the expansion of spacetime. The second part of the statement tells us that, in turn, the function $a(t)$ is determined by the matter, or more precisely the energy density, in the universe. Here we will first describe the kind of substances that fill the universe and then, in Section 1.2.3 turn to their effect on the expansion.

1.2.1 Perfect Fluids

The cosmological principle guides us to model the contents of the universe as a homogeneous and isotropic fluid. The lumpy, clumpy nature of galaxies that we naively observe is simply a consequence of our small perspective. Viewed from afar, we should think of these galaxies as like atoms in a cosmological fluid. Moreover, as we will learn, the observable galaxies are far from the most dominant energy source in the universe.

We treat all such sources as homogeneous and isotropic perfect fluids. This means that they are characterised by two quantities: the *energy density* $\rho(t)$ and the *pressure* $P(t)$. (If you've taken a course in fluid mechanics, you will be more used to thinking of $\rho(t)$ as the mass density. In the cosmological, or relativistic context, this becomes the total energy density.)

The Equation of State

For any fluid, there is a relation between the energy and pressure, $P = P(\rho)$, known as the *equation of state*.

We will need the equation of state for two, different kinds of fluids. Both of these fluids contain constituent "atoms" of mass m which obey the relativistic energy-momentum relation

$$E^2 = p^2c^2 + m^2c^4 \tag{1.32}$$

The two fluids come from considering this equation in two different regimes:

- Non-Relativistic Limit: $pc \ll mc^2$. Here the energy is dominated by the mass, $E \approx mc^2$, and the velocity of the atoms is $\mathbf{v} \approx \mathbf{p}/m$.
- Relativistic Limit: $pc \gg mc^2$. Now the energy is dominated by the momentum, $E \approx pc$, and the velocity of the atoms approaches the speed of light $|\mathbf{v}| \approx c$.

Suppose that there are N such atoms in a volume V. In general, these atoms will not have a fixed momentum and energy, but instead the number density $n(p)$ will be some distribution. Because the fluid is isotropic, this distribution can depend only on the magnitude of momentum $p = |\mathbf{p}|$. It is normalised by

$$\frac{N}{V} = \int_0^\infty dp\ n(p)$$

The pressure of a gas is defined to be force per unit area. For our purposes, a better definition is the flux of momentum across a surface of unit area. This is equivalent

to the earlier definition because, if the surface is a solid wall, the momentum must be reflected by the wall resulting in a force. However, the "flux" definition can be used anywhere in the fluid, not just at the boundary where there's a wall. Because the fluid is isotropic, we are free to choose this area to be the (x, y)-plane. Then, we have

$$P = \int_0^\infty dp\ v_z p_z n(p)$$

(If this is unfamiliar, an elementary derivation of this formula is given later in Section 2.1.2.) Because $\mathbf{v}$ and $\mathbf{p}$ are parallel, we can write

$$\mathbf{v} \cdot \mathbf{p} = vp = v_x p_x + v_y p_y + v_z p_z = 3 v_z p_z$$

where the final equality is ensured by isotropy. This then gives us

$$P = \frac{1}{3} \int_0^\infty dp\ vp\, n(p) \tag{1.33}$$

Now we can relate this to the energy density in the two cases. First, the non-relativistic gas. In this case, $p \approx mv$ so we have

$$P_{\rm non-rel} \approx \frac{1}{3} \int_0^\infty dp\ mv^2 n(p) = \frac{1}{3} \frac{N}{V} m \langle v^2 \rangle \tag{1.34}$$

where $\langle v^2 \rangle$ is the average square-velocity in the gas.

For cosmological purposes, our interest is in the total energy (1.32) and this is dominated by the contribution from the mass $E \approx mc^2 + \ldots$. If we relate the pressure of a non-relativistic gas to this total energy E, we have

$$P_{\rm non-rel} = \frac{NE}{V} \frac{\langle v^2 \rangle}{c^2}$$

Since $\langle v^2 \rangle / c^2 \ll 1$, we say that the pressure of a non-relativistic gas is simply

$$P_{\rm non-rel} \approx 0$$

Note that this is the same pressure that keeps balloons afloat and your eardrums healthy: it's not really vanishing. But it is negligible when it comes to its effect on the expansion of the universe. (We will, in fact, revisit this in Section 2 where we'll see that the pressure does give rise to important phenomena in the early universe.)

Cosmologists refer to a non-relativistic gas as *dust*, a name designed to reflect the fact that it just hangs around and is boring. Examples of dust include galaxies, dark matter, and hydrogen atoms floating around and not doing much. We will also refer to dust simply as *matter*.

We can repeat this for a gas of relativistic particles with $v \approx c$ and $E \approx pc$. Now the formula for the pressure (1.33) becomes

$$P_{\rm rel} \approx \frac{1}{3}\int_0^\infty dp\ vp\, n(p) \approx \frac{1}{3}\int_0^\infty dp\ E\, n(p) = \frac{N\langle E\rangle}{3V}$$

with $\langle E\rangle$ the average energy of a particle. The energy density is $\rho = N\langle E\rangle/V$, so the relativistic gas obeys the equation of state

$$P_{\rm rel} = \frac{1}{3}\rho$$

Cosmologists refer to such a relativistic gas as *radiation*. Examples of radiation include the gas of photons known as the cosmic microwave background, gravitational waves, and neutrinos.

Most of the equations of state we meet in cosmology have the simple form

$$P = w\rho \tag{1.35}$$

for some constant w. As we have seen, dust has $w = 0$ and radiation has $w = 1/3$. We will meet other, more exotic fluids as the course progresses.

There is an important restriction on the equation of state. The speed of sound c_s in a fluid is given by

$$c_s^2 = c^2\,\frac{dP}{d\rho}$$

We will derive this formula in Section 3.1.1, but for now we simply quote it. It's important that the speed of sound is less than the speed of light. (Remember: nothing can beat light in a race.) This means that to be consistent with relativity, we must have $w \leq 1$. In fact, the more exotic substances we will meet will have $w < 0$, suggesting an imaginary sound speed. What this is really telling us is that substances with $w < 0$ do not support propagating sound waves, with perturbations decaying exponentially in time.

An Aside: The Equation of State and Temperature

In many other areas of physics, the equation of state is usually written in terms of the temperature T of a fluid. For example, the ideal gas equation relates the pressure P and volume V as

$$PV = Nk_BT \tag{1.36}$$

where N is the number of particles and k_B the Boltzmann constant. (You may have seen this written in chemist's notation $Nk_B = nR$ where n is the number of moles and R the gas constant. Our way is better.) The equations of state that we're interested in can be viewed in this way if we relate T/V to the energy density.

For example, starting from our expression, in (1.34) we derived an expression for the pressure of a non-relativistic gas: $P_{\rm non-rel} \approx Nm\langle v^2\rangle/3V$. This coincides with the ideal gas law if we relate the temperature to the average kinetic energy of an atom in the gas through

$$\frac{1}{2}m\langle v^2\rangle = \frac{3}{2}k_B T \tag{1.37}$$

We will revisit this in Section 2.1 and gain a better understanding of this result and the role played by temperature.

1.2.2 The Continuity Equation

As the universe expands, we expect the energy density (of any sensible fluid) to dilute. The way this happens is dictated by the conservation of energy, also known as the continuity equation.

A proper discussion of the continuity equation requires the machinery of general relativity. This is one of a number of places were we will revert to some simple Newtonian thinking to derive the correct equation. Such derivations are not entirely convincing, not least because it's unclear why they would be valid when applied to the entire universe. Nonetheless, they will give the correct answer. A more rigorous approach can be found in the lectures on General Relativity.

Consider a gas trapped in a box of volume V. The gas exerts pressure on the sides of the box. If the box increases in size, as shown in the figure, then the change of volume is $dV = \text{Area} \times dx$. The work done by the gas is Force $\times\, dx = (PA)dx = p\,dV$, and this reduces the internal energy of the gas. We have

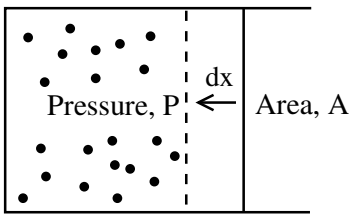

Figure 12:

$$dE = -P\,dV$$

This is a simple form of the *first law of thermodynamics*, valid for reversible or adiabatic processes. It is far from obvious that we can view the universe as a box filled with gas and naively apply this formula. Nonetheless, it happily turns out that the final result

agrees with the more rigorous GR approach so we will push ahead, and invoke the time dependent version of the first law,

$$\frac{dE}{dt} = -P\frac{dV}{dt} \tag{1.38}$$

Now consider a small region of fluid, in co-moving volume V_0. The physical volume is

$$V(t) = a^3(t)V_0 \quad \Rightarrow \quad \frac{dV}{dt} = 3a^2\dot{a}V_0$$

Meanwhile, the energy in this volume is

$$E = \rho a^3 V_0 \quad \Rightarrow \quad \frac{dE}{dt} = \dot{\rho}a^3V_0 + 3\rho a^2\dot{a}V_0$$

The first law (1.38) then becomes

$$\dot{\rho} + 3H\left(\rho + P\right) = 0 \tag{1.39}$$

This slightly unfamiliar equation is the expression of energy conservation in a cosmological setting.

Before we proceed, a warning: energy is a famously slippery concept in general relativity, and we will meet things later which, taken naively, would seem to violate energy conservation. For example, in Section 1.3.3, we will meet a fluid with equation of state $\rho = -P$. For such a fluid, $\dot{\rho} = 0$ which means that the energy density remains constant even as the universe expands. Such is the way of the world and we need to get used to it. If this makes you nervous, recall that the usual derivation of energy conservation, via Noether's theorem, holds only in time independent settings. So perhaps it's not so surprising that energy conservation takes a somewhat different form in an expanding universe.

If we specify an equation of state $P = w\rho$, as in (1.35), then we can integrate the continuity equation (1.39) to determine how the energy density depends on the scale factor. We have

$$\begin{aligned} \frac{\dot{\rho}}{\rho} = -3(1+w)\frac{\dot{a}}{a} \quad &\Rightarrow \quad \log(\rho/\rho_0) = -3(1+w)\log\, a \\ &\Rightarrow \quad \rho(t) = \rho_0 a^{-3(1+w)} \end{aligned} \tag{1.40}$$

with $\rho_0 = \rho(t_0)$ and we've used the fact that $a(t_0) = 1$.

We can look at how this behaves in simple examples. For dust (also known as matter), we have $w = 0$ and so

$$\rho_m \sim \frac{1}{a^3}$$

This makes sense. As the universe expands, the volume increases as a^3, and so the energy density decreases as $1/a^3$.

For radiation, we instead find

$$\rho_r \sim \frac{1}{a^4} \tag{1.41}$$

This also makes sense. The energy density is diluted as $1/a^3$ but, on top of this, there is also a redshift effect which shifts the frequency, and hence the energy, by a further power of $1/a$.

The fact that the energy densities of dust and radiation scale differently plays a crucial role in our cosmological history. As we shall see in Section 1.4, our current universe has much greater energy density in dust than in radiation. However, this wasn't always the case. There was a time in far past when the converse was true, with the radiation subsequently diluting away faster. We'll see other contributions to the energy density of the universe that have yet different behaviour.

1.2.3 The Friedmann Equation

> "Friedmann more than once said that his task was to indicate the possible solutions of Einstein's equations, and that the physicists could do what they wished with these solutions"
>
> Vladimir Fock, on his friend Alexander Friedmann

Finally we come to the main part of the story: we would like to describe how the perfect fluids which fill all of space affect the expansion of the universe. We start by giving the answer. The dynamics of the scale factor is dictated by the energy density $\rho(t)$ through the *Friedmann equation*

$$H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3c^2}\rho - \frac{kc^2}{R^2a^2} \tag{1.42}$$

Here R is some fixed scale, as in the FRW metric (1.12), $k = -1, 0, +1$ determines the curvature of space, and G is Newton's gravitational constant

$$G \approx 6.67 \times 10^{-11}\ \mathrm{m^3\,kg^{-1}\,s^{-2}}$$

The Friedmann equation is arguably the most important equation in all of cosmology. Taken together with the continuity equation (1.39) and the equation of state (1.35), they provide a closed system which can be solved to determine the history and fate of the universe itself.

At this point, I have a confession to make. The only honest derivation of the Friedmann equation is in the framework of General Relativity Here we can only present a dishonest derivation, using Newtonian ideas. In an attempt to alleviate the shame, I will at least be open about where the arguments are at their weakest.

First, we work in flat space, with $k = 0$. This, of course, is the natural habitat for Newtonian gravity. Nonetheless, we will see the possibility of a curvature term $-k/a^2$ in the Friedmann equation, re-emerging at the end of our derivation.

Our discussions so far prompt us to consider an infinite universe, filled with a constant matter density. That, it turns out, is rather subtle in a Newtonian setting. Instead, we consider a ball of uniform density of size L, expanding outwards away from the origin, and subsequently pretend that we can take $L \to \infty$.

Consider a particle (or element of fluid) of mass m at some position $\mathbf{x}$ with $r = |\mathbf{x}| \ll L$. It will experience the force of gravity in the form of Newton's inverse-square law. But a rather special property of this law states that, for a spherically symmetric distribution of masses, the gravitational force at some point $\mathbf{x}$ depends only on the masses at distances smaller than r and, moreover, acts as if all the mass is concentrated at the origin.

This statement is simplest to prove if we formulate the gravitational force law as a kind of Gauss' law,

$$\mathbf{F}_{\rm grav} = -m\nabla\Phi \quad \text{where} \quad \nabla^2\Phi = \frac{4\pi G}{c^2}\rho$$

with Φ the gravitational potential. The (perhaps) unfamiliar factor of c^2 in the final equation arises because, for us, ρ is the energy density, rather than mass density. We then integrate both sides over a ball V of radius x, centred at the origin. Using the kind of symmetry arguments that we used extensively in the lectures on Electromagnetism, we have

$$\int_S \nabla\Phi \cdot d\mathbf{S} = \int_V \frac{4\pi G}{c^2}\rho\, dV \quad \Rightarrow \quad \nabla\Phi(r) = \frac{GM(r)}{r^2}$$

where $M(r) = 4\pi\rho r^3/3c^2$ is the mass contained inside the ball of radius r. This means that the acceleration of the particle at $\mathbf{x}$ is given by

$$m\ddot{r} = -\frac{GmM(r)}{r^2}$$

We multiply by $\dot{r}$ and integrate. As the ball expands with $\dot{r} \neq 0$, the total mass contained with a ball of radius $r(t)$ does not change, so $\dot{M} = 0$. We then get

$$\frac{1}{2}\dot{r}^2 - \frac{GM(r)}{r} = E \tag{1.43}$$

where we recognise E as the energy (per unit mass) of the particle. Finally, we describe the position $\mathbf{x}$ of the particle in a way that chimes with our previous cosmological discussion, introducing a scale factor $a(t)$

$$\mathbf{x}(t) = a(t)\mathbf{x}_0$$

Substituting this into (1.43) and rearranging gives

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3c^2}\rho - \frac{C}{a^2} \tag{1.44}$$

where $C = -2E/|\mathbf{x}_0|^2$ is a constant. This is remarkably close to the Friedmann equation (1.42). The only remaining issue is why we should identify the constant C with the curvature kc^2/R^2. There is no good argument here and, indeed, we shouldn't expect one given that the whole Newtonian derivation took place in a flat space. It is, unfortunately, simply something that you have to suck up.

There is, however, an analogy which makes the identification $C \sim k$ marginally more palatable. Recall that a particle has reached escape velocity if its total energy $E > 0$. Conversely, if $E < 0$, the particle comes crashing back down. For us, the case of $E < 0$ means $C > 0$ which, in turn, corresponds to positive curvature. We will see in Section 1.3.2 that a universe with positive curvature will, under many circumstances, ultimately suffer a big crunch. In contrast, a negatively curved space $k < 0$ will keep expanding forever.

Clearly the derivation above is far from rigorous. There are at least two aspects that should give us pause. First, when we assumed $\dot{M} = 0$, we were implicitly restricting ourselves to non-relativistic matter with $\rho \sim 1/a^3$. It turns out that in general relativity, the Friedmann equation also holds for any other scaling (1.40) of ρ.

However, the part of the above story that should make you feel most queasy is replacing an infinitely expanding universe, with an expanding ball of finite size L. This introduces an origin into the story, and gives a very misleading impression of what the expansion of the universe means. In particular, if we dial the clock back to $a(t) = 0$ in this scenario, then all matter sits at the origin. This is one of the most popular misconceptions about the Big Bang and it is deeply unfortunate that it is reinforced by the derivation above. Nonetheless, the arguments that lead to (1.44) do provide some physical insight into the meaning of the various terms that can be hard to extract from the more formal derivation using general relativity. So let us wash the distaste from our mouths, and proceed with understanding the universe.

1.3 Cosmological Solutions

We now have a closed set of equations that describe the evolution of the universe. These are the Friedmann equation,

$$H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3c^2}\rho - \frac{kc^2}{R^2a^2} \tag{1.45}$$

the continuity equation,

$$\dot{\rho} + 3H\left(\rho + P\right) = 0$$

and the equation of state

$$P = w\rho$$

In this section, we will solve them. Our initial interest will be on a number of designer universes whose solutions are particularly simple. Then, in Section 1.4, we describe the solutions of relevance to our universe.

1.3.1 Simple Solutions

To solve the Friedmann equation, we first need to decide what fluids live in our universe. In general, there will be several different fluids. If they share the same equation of state (e.g. dark matter and visible matter) then we can, for cosmological purposes, just treat them as one. However, if the universe contains fluids with different equations of state, we must include them all. In this case, we write

$$\rho = \sum_w \rho_w$$

As we have seen in (1.40), each component scales independently as

$$\rho_w = \frac{\rho_{w,0}}{a^{3(1+w)}} \tag{1.46}$$

where $\rho_{w,0} = \rho_w(t_0)$. Substituting this into the Friedmann equation then leaves us with a tricky-looking non-linear differential equation for a.

Life is considerably simpler if we restrict attention to a flat $k = 0$ universe with just a single fluid component. In this case, using (1.46), we have

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{D^2}{a^{3(1+w)}} \tag{1.47}$$

where $D^2 = 8\pi G\rho_{w,0}/3c^2$ is a constant. The solution is

$$a(t) = \left(\frac{t}{t_0}\right)^{2/(3+3w)} \tag{1.48}$$

The various constants have been massaged into $t_0 = (\frac{3}{2}(1+w)D)^{-1}$ so that we recover our convention $a_0 = a(t_0) = 1$. There is also an integration constant which we have set to zero. This corresponds to picking the time of the Big Bang, defined by $a(t_{BB}) = 0$ to be $t_{BB} = 0$. With this choice, t_0 is identified with the age of the universe.

Let's look at this solution in a number of important cases

- Dust ($w = 0$): For a flat universe filled with dust-like matter (i.e. galaxies, or cold dark matter), we have

  $$a(t) = \left(\frac{t}{t_0}\right)^{2/3} \tag{1.49}$$

  This is known as the *Einstein-de Sitter universe* (not to be confused with either the Einstein universe or the de Sitter universe, both of which we shall meet in Section 1.3.3). The exponent 2/3 is the same 2/3 that appears in Kepler's third law: the radius R of a planet's orbit is related to its period by $R \sim T^{2/3}$. Both follow by simple dimensional analysis in Newtonian gravity.

  The Hubble constant is

  $$H_0 = \frac{2}{3}\frac{1}{t_0}$$

  If we lived in such a place, then a measurement of H_0 would immediately tell us the age of the universe $t_0 = \frac{2}{3}H_0^{-1}$. Using the observed value of $H_0 \approx 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ gives

  $$t_0 \approx 9 \times 10^9 \text{ years} \tag{1.50}$$

The extra factor of 2/3 brings us down from the earlier estimate of 14 billion years in (1.23) to 9 billion years. This is problematic since there are stars in the universe that appear to be older than this.

Finally note that in the Einstein-de Sitter universe the matter density scales as

$$\rho(t) = \frac{c^2}{6\pi G}\frac{1}{t^2} \tag{1.51}$$

In particular, there is a direct relationship between the age of the universe and the present day matter density. We'll revisit this relationship later.

- Radiation ($w = 1/3$): For a flat universe filled with radiation (e.g. light), we have

$$a(t) = \left(\frac{t}{t_0}\right)^{1/2}$$

Once again, there is a direct relation between the Hubble constant and the age of the universe, now given by $t_0 = \frac{1}{2}H_0^{-1}$. In a radiation dominated universe, the energy density scales as

$$\rho(t) = \frac{3c^2}{32\pi G}\frac{1}{t^2}$$

- Curvature ($w = -1/3$): We can also apply the calculation above to a universe with curvature a term, which is devoid of any matter. Indeed, the curvature term in (1.45) acts just like a fluid (1.46) with $w = -1/3$. In the absence of any further fluid contributions, the Friedmann equation only has solutions for a negatively curved universe, with $k = -1$. In this case,

$$a(t) = \frac{t}{t_0}$$

This is known as the *Milne universe*.

A Comment on Multi-Component Solutions

If the universe has more than one type of fluid (or a fluid and some curvature) then it is more tricky to write down analytic solutions to the Friedmann equations. Nonetheless, we can build intuition for these solutions using our results above, together with the observation that different fluids dilute away at different rates. For example, we have seen that

$$\rho_m \sim \frac{1}{a^3} \quad \text{and} \quad \rho_r \sim \frac{1}{a^4}$$

This means means that, in a universe with both dust and radiation (like the one we call home) there will be a period in the past, when a is suitably small, when we necessarily have $\rho_r \gg \rho_m$. As a increases there will be a time when the energy density of the two are roughly comparable, before we go over to another era with $\rho_m \gg \rho_r$. In this way, the history of the universe is divided into different epochs. When one form of energy density dominates over the other, the expansion of the universe is well-approximated by the single-component solutions we met above .

The Big Bang Revisited: A Baby Singularity Theorem

All of the solutions we met above have a Big Bang, where $a = 0$. It is natural to ask: is this a generic feature of the Friedmann equation with arbitrary matter and curvature?

Within the larger framework of general relativity, there are a number of important theorems which state that, under certain circumstances, singularities in the metric necessarily arise. The original theorems, due to Penrose (for black holes) and Hawking (for the Big Bang), are tour-de-force pieces of mathematical physics. You can learn about them next year. Here we present a simple Mickey mouse version of the singularity theorem for the Friedmann equation.

We start with the Friedmann equation, written as

$$\dot{a}^2 = \frac{8\pi G}{3c^2}\rho a^2 - \frac{kc^2}{R^2}$$

Differentiating both sides with respect to time gives

$$2\ddot{a}\dot{a} = \frac{8\pi G}{3c^2}\left(\dot{\rho}a^2 + 2\rho\dot{a}a\right) = \frac{8\pi G}{3c^2}\left(-3\dot{a}a(\rho + P) + 2\rho\dot{a}a\right)$$

where, in the second equality, we have used the continuity equation $\dot{\rho} + 3H(\rho + P) = 0$ Rearranging gives the *acceleration equation*

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2}(\rho + 3P) \tag{1.52}$$

This is also known as the *Raychaudhuri equation* and will be useful in a number of places in this course. (It is a special case of the real Raychaudhuri equation, which has application beyond cosmology.) Using this result, we can prove the following:

Claim: If matter obeys the *strong energy condition*

$$\rho + 3P \geq 0 \tag{1.53}$$

then there was a singularity at a finite time t_{BB} in the past where $a(t_{BB}) = 0$. Furthermore, $t_0 - t_{BB} \leq H_0^{-1}$.

Proof: The strong energy condition immediately tells us that $\ddot{a}/a \leq 0$. This is the statement that the universe is decelerating, meaning that it must have been expanding faster in the past.

Suppose first that $\ddot{a} = 0$. In this case we must have $a(t) = H_0 t + \text{const}$. (We have used the fact that $H_0 = \dot{a}_0$ since $a_0 = 1$). This is the dotted line shown in the figure. If this is the case, the Big Bang occurs at $t_0 - t_{BB} = H_0^{-1}$. But the strong energy condition ensures that $\ddot{a} \leq 0$, so the dotted line in the figure provides an upper bound on the scale factor. In such a universe, the Big Bang must occur at $t_0 - t_{BB} \leq H_0^{-1}$. □

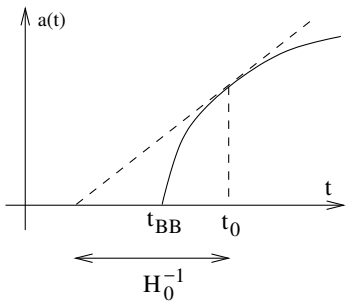

Figure 13:

The proof above is so simple because we have restricted attention to the homogeneous and isotropic FRW universe. Hawking's singularity theorem (proven in his PhD thesis) shows the necessity of a singularity even in the absence of such assumptions.

The strong energy condition is obeyed by all conventional matter, including dust and radiation. However, it's not hard to find substances which violate it, and we shall meet examples as we go along. When the strong energy condition is violated, we have an accelerating universe with $\ddot{a} > 0$. In this case, the single component solutions (1.48) still have a Big Bang singularity. However, the argument above cannot rule out the possibility of more complicated solutions which avoid this.

The Future Revisited: Cosmological Event Horizons

Recall from section 1.1.4 the idea of an event horizon: for certain universes, it may be that our friends in distant galaxies get swept away from us by the expansion of space and are lost to us forever. At a time t, the furthest distance with which we can communicate, $r_{\rm max}$ is governed by the equation (1.25)

$$c\int_t^\infty \frac{dt'}{a(t')} = \int_0^{r_{\rm max}(t)} \frac{dr}{\sqrt{1-kr^2/R^2}}$$

If the integral on the left converges then $r_{\rm max}$ is finite and there is a cosmological horizon.

When does this happen? If the late time universe is dominated by a single component with expansion given by $a \sim t^{2/(3+3w)}$ as in (1.48) then

$$\int \frac{dt}{a(t)} \sim \int \frac{dt}{t^{2/(3+3w)}} \sim t^{(3w+1)/(3w+3)}$$

For $w \geq -1/3$, the integral diverges and there is no event horizon. (In the limiting case of $w = -1/3$, the integral is replaced by $\log t$.) For $-1 \leq w < -1/3$, the integral converges and there is a horizon.

Fluids with $w < -1/3$ are precisely those which violate the strong energy condition (1.53). We learn that cosmological event horizons arise whenever the late time expansion of the universe is accelerating, rather than decelerating.

1.3.2 Curvature and the Fate of the Universe

Let's look again at a flat universe, with $k = 0$. The Friedmann equation (1.45) tells us that for such a universe to exist, something rather special has to happen, because the energy density of the universe today ρ_0 has to be precisely correlated with the Hubble constant

$$H_0^2 = \frac{8\pi G}{3c^2}\rho_0$$

We saw such behaviour in our earlier solutions. For example, this led us to the result (1.51) which relates the energy density of an Einstein-de Sitter universe to the current age of the universe.

In principle, this gives a straightforward way to test whether the universe is flat. First, you measure the expansion rate as seen in H_0. Then you add up all the energy in the universe and see if they match. In practice, this isn't possible because, as we shall see, much of the energy in the universe is invisible.

What happens if we have a universe with some small curvature and, say, a large amount of conventional matter with $w = 0$? We can think of the curvature term in the Friedmann equation as simply another contribution to the energy density, ρ_k, one which dilutes away more slowly that the matter contribution,

$$\rho_m \sim \frac{1}{a^3} \quad \text{and} \quad \rho_k \sim \frac{1}{a^2}$$

This tells us that, regardless of their initial values, if we wait long enough then the curvature of space will eventually come to dominate the dynamics.

If we start with $\rho_m > \rho_k$, then there will be a moment when the two are equal, meaning

$$\frac{8\pi G}{3c^2}\rho_m = \frac{|k|c^2}{R^2a^2}$$

For a negatively curved universe, with $k = -1$, the Friedmann equation (1.45) gives $\dot{a} > 0$. However, for a positively curved universe, with $k = +1$, we find $\dot{a} = 0$ at the moment of equality. In other words, the universe stops expanding. In fact, as we now see, such a positively curved universe subsequently contracts until it hits a big crunch.

Perhaps surprisingly, it is possible to find an exact solution to the Friedmann equation with both matter and curvature. To do this, it is useful to work in conformal time (1.26), defined by

$$\tau(t) = \int_0^t \frac{dt'}{a(t')} \quad \Rightarrow \quad \frac{d\tau}{dt} = \frac{1}{a} \tag{1.54}$$

We further define the dimensionless time coordinate $\tilde{\tau} = c\tau/R$. (In flat space, with $k = 0$, just pick a choice for R; it will drop out in what follows.) Finally, we define

$$h = \frac{a'}{a} \quad \text{with} \quad a' = \frac{da}{d\tilde{\tau}}$$

In these variables, one can check that the Friedmann equation (1.45) becomes

$$h^2 + k = \frac{8\pi G R^2}{3c^4}\rho a^2 \tag{1.55}$$

Rather than solve this in conjunction with the continuity equation, it turns out to be more straightforward to look at the acceleration equation (1.52). A little algebra shows that, for matter with $P = 0$, the acceleration equation becomes

$$h' = -\frac{4\pi G R^2}{3c^4}\rho a^2 \quad \Rightarrow \quad 2h' + h^2 + k = 0 \tag{1.56}$$

where, to get the second equation, we have simply used (1.55). Happily this latter equation is independent of ρ and we can go ahead and solve it. The solutions are:

$$h(\tilde{\tau}) = \begin{cases} \cot(\tilde{\tau}/2) & k = +1 \\ 2/\tilde{\tau} & k = 0 \\ \coth(\tilde{\tau}/2) & k = -1 \end{cases}$$

We can then solve $h = a'/a$ to derive an expression for the scale factor $a(\tilde{\tau})$ as a function of $\tilde{\tau}$,

$$a(\tilde{\tau}) = A \times \begin{cases} \sin^2(\tilde{\tau}/2) & k = +1 \\ \tilde{\tau}^2 & k = 0 \\ \sinh^2(\tilde{\tau}/2) & k = -1 \end{cases} \tag{1.57}$$

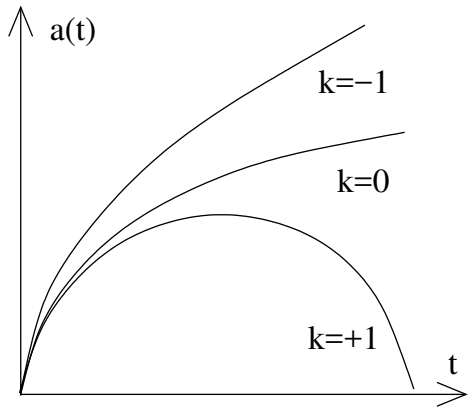

Figure 14: The FRW scale factor for a matter dominated universe with curvature.

with A an integration constant. We see that, as advertised, the positively curved $k = 1$ universe eventually re-collapses, with the Big Crunch occurring at conformal time $\tau = 2\pi R/c$. In contrast, the negatively curved $k = -1$ universe expands for ever. The flat space $k = 0$ separates these two behaviours.

Finally, we can use the solution for the scale factor to determine how conformal time (1.54) scales with our original time coordinate t,

$$t = \frac{RA}{2c} \times \begin{cases} \tilde{\tau} - \sin\tilde{\tau} & k = +1 \\ \tilde{\tau}^3 & k = 0 \\ \sinh\tilde{\tau} - \tilde{\tau} & k = -1 \end{cases} \tag{1.58}$$

In the $k = 0$ case, this reproduces our previous result (1.49) for the expansion of the Einstein-de Sitter universe. The resulting scale factors $a(t)$ are sketched in Figure 14.

There are a couple of lessons to take from this calculation. The first is that a flat universe is dynamically unstable, rather like a pencil balancing on its tip. Any small initial curvature will grow and dominate the late time behaviour.

The second lesson comes with an important caveat. The result above suggests that a measurement of curvature of the space will tell us the ultimate fate of the universe. If we find $k = 1$, then we are doomed to suffer a Big Crunch. On the other hand, a curvature of $k = -1$ or $k = 0$ means that universe expands for ever, becoming increasingly desolate and lonely. However, this conclusion relies on the assumption that the dominant energy in the universe is matter. In fact, it's not hard to show that the conclusion is unaltered provided that all energies in the universe dilute away faster

than the curvature. However, as we will now see, there are more exotic fluids at play in the universe for which the conclusion does not hold.

1.3.3 The Cosmological Constant

The final entry in the dictionary of cosmological fluids is both the most strange and, in some ways, the most natural. A *cosmological constant* is a fluid with equation of state $w = -1$. The associated energy density is denoted ρ_Λ and obeys

$$\rho_\Lambda = -P$$

First the strange. The continuity equation (1.39) tells us that such an energy density remains constant over time: $\rho_\Lambda \sim a^0$. Naively, that would seem to violate the conservation of energy. However, as stressed previously, energy is a rather slippery concept in an expanding universe and the only thing that we have to worry about is the continuity equation (1.39) which is happily obeyed. So this is something we will just have to live with. For now, note that any universe with $\rho_\Lambda \neq 0$ will ultimately become dominated by the cosmological constant, as all other energy sources dilute away.

Now the natural. The cosmological constant is something that you've seen before. Recall that whenever you write down the energy of a system, any overall constant shift of the energy is unimportant and does not affect the physics. For example, in classical mechanics if we have a potential $V(\mathbf{x})$, then the force is $\mathbf{F} = -\nabla V$ which cares nothing about the constant term in V. Similarly, in quantum mechanics we work with the Hamiltonian H, and adding an overall constant is irrelevant for the physics. However, when we get to general relativity, it becomes time to pay the piper. In the context of general relativity, all energy gravitates, including the constant energy that we previously neglected. And the way this constant manifests itself is as a cosmological constant. For this reason, the cosmological constant is also referred to as *vacuum energy*.

Strictly speaking, ρ_Λ is the vacuum energy density, while the cosmological constant Λ is defined as

$$\rho_\Lambda = \frac{\Lambda c^2}{8\pi G}$$

so Λ has dimensions of (time)$^{-2}$. (Usually, by the time people get to describing the cosmological constant, they have long set $c = 1$, so other definitions may differ by hidden factors of c.) Here we will treat the terms "cosmological constant" and "vacuum

energy" as synonymous. In the presence of a cosmological constant and other matter, the Friedmann equation becomes

$$H^2 = \frac{8\pi G}{3c^2}\rho + \frac{\Lambda}{3} - \frac{kc^2}{R^2a^2} \tag{1.59}$$

We now solve this in various cases.

de Sitter Space

First, consider a universe with positive cosmological constant $\Lambda > 0$. If we empty it of all other matter, so that $\rho = 0$, then we can solve the Friedmann equation for any choice of curvature $k = -1, 0, +1$ to give

$$a(t) = \begin{cases} A\cosh\left(\sqrt{\Lambda/3}\,t\right) & k = +1 \\ A\exp\left(\sqrt{\Lambda/3}\,t\right) & k = 0 \\ A\sinh\left(\sqrt{\Lambda/3}\,t\right) & k = -1 \end{cases}$$

where $A^2 = 3c^2/\Lambda R^2$ for the $k = \pm 1$ solutions, and is arbitrary for the $k = 0$ solution. At large time, all of these solutions exhibit exponential behaviour, independent of the spatial curvature. In fact, it turns out (although we won't show it here) that each of these solutions describes the same spacetime, but with different coordinates that slice spacetime into space+time in different ways . This spacetime is known as *de Sitter space*.

The $k = +1$ solution most accurately represents the geometry of de Sitter space because it uses coordinates which cover the whole spacetime . It shows a contracting phase when $t < 0$, followed by a phase of accelerating expansion when $t > 0$. Crucially, there is no Big Bang when $a = 0$. In contrast, the $k = 0$ and $k = -1$ coordinates give a slightly misleading view of the space, because they suggest a Big Bang when $t = -\infty$ and $t = 0$ respectively. You need to work harder to show that actually this is an artefact of the choice of coordinates (a so-called "coordinate singularity") rather than anything physical. These kind of issues will be addressed in next term's course on general relativity.

To better understand this spacetime and, in particular, the existence of cosmological horizons, it is best to work with $k = +1$ and conformal time, $\tau \in (-\pi/2, +\pi/2)$, given by

$$\cos\left(\sqrt{\Lambda/3}\tau\right) = \left[\cosh\left(\sqrt{\Lambda/3}t\right)\right]^{-1}$$

You can check that $d\tau/dt = 1/\cosh^2(\sqrt{\Lambda/3}\,t)$, which, up to an overall unimportant scale, is the definition of conformal time (1.26). In these coordinates, the metric for de Sitter space becomes

$$ds^2 = \frac{1}{\cos^2(\sqrt{\Lambda/3}\tau)}\left[-c^2d\tau^2 + R^2d\chi^2 + R^2\sin^2\chi(d\theta^2 + \sin^2\theta\, d\phi^2)\right]$$

where we're using the polar coordinates (1.6) on the spatial $\mathbf{S}^3$. We now consider a fixed θ and ϕ and draw the remaining 2d spacetime in the $(c\tau, \chi)$ plane where $\tau \in (-\pi/2, \pi/2)$ and $\chi \in [0, \pi]$. The left-hand edge of the diagram can be viewed as the north pole of $\mathbf{S}^3$, $\chi = 0$, while the right-hand edge of the diagram is the south pole $\chi = \pi$. The purpose of this diagram is not to exhibit distances between points, because these are distorted by the $1/\cos^2\tau$ factor in front of the metric. Instead, the diagram shows only the causal structure, with 45° lines denoting light rays.

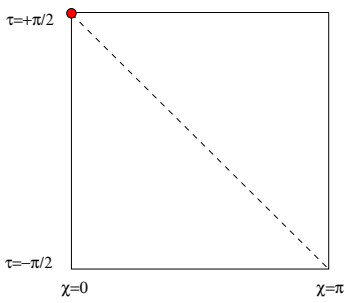

Figure 15:

Consider an observer sitting at the north pole. She has a particle horizon and an event horizon. Even if she waits forever, as shown in the figure, there will be part of the spacetime that she never sees.

Anti-de Sitter Space

We could also look at solutions with $\Lambda < 0$, again devoid of any matter so $\rho = 0$. A glance at the Friedmann equation (1.59) shows that such solutions can only exist when $k = -1$. In this case, the scale factor is given by

$$a(t) = A\sin\left(\sqrt{-\Lambda/3}\,t\right)$$

This is known as *anti-de Sitter space*. It has, as far as we can tell, no role to play in cosmology. However it has become rather important as a testing ground for ideas in quantum gravity and holography. We will not discuss it further here.

Matter + Cosmological Constant

For a flat $k = 0$ universe, we can find a solution for a positive cosmological constant $\Lambda > 0$, with matter $\rho_m \sim 1/a^3$. We write the Friedmann equation as

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3c^2}\left(\rho_\Lambda + \frac{\rho_0}{a^3}\right)$$

This has the solution

$$a(t) = \left(\frac{\rho_0}{\rho_\Lambda}\right)^{1/3} \sinh^{2/3}\left(\frac{\sqrt{3\Lambda}t}{2}\right) \tag{1.60}$$

There are a number of comments to make about this. First note that, in contrast to de Sitter space, the Big Bang has unavoidably reappeared in this solution at $t = 0$ where $a(t = 0) = 0$. This, it turns out, is generic: any universe more complicated than de Sitter (like ours) has a Big Bang singularity.

The present day time t_0 is defined, as always, by $a(t_0) = 1$. There is also another interesting time, $t_{\rm eq}$, where we have matter-vacuum energy equality, so that $\rho_\Lambda = \rho_0/a^3$. This occurs when

$$\sinh\left(\frac{\sqrt{3\Lambda}t_{\rm eq}}{2}\right) = 1 \tag{1.61}$$

At late times, the solution (1.60) coincides with the de Sitter expansion $a(t) \sim e^{\sqrt{\Lambda/3}t}$, telling us that the cosmological constant is dominating as expected. Meanwhile, at early times we have $a \sim t^{2/3}$ and we reproduce the characteristic expansion of the Einstein-de Sitter universe (1.49).

An Historical Curiosity: The Einstein Static Universe

The cosmological constant was first introduced by Einstein in 1917 in an attempt to construct a static cosmology. This was over a decade before Hubble's discovery of the expanding universe.

The acceleration equation (1.52)

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2}(\rho + 3P) \tag{1.62}$$

tells us that a static universe is only possible if $\rho = -3P$. Obviously this is not possible if we have only matter ρ_m with $P_m = 0$ or only a cosmological constant $\rho_\Lambda = -P_\Lambda$. But in a universe with both, we can have

$$\rho = \rho_m + \rho_\Lambda = -3P = 3\rho_\Lambda \quad \Rightarrow \quad \rho_m = 2\rho_\Lambda$$

The Friedmann equation (1.59) is then

$$H^2 = \frac{8\pi G}{3c^2}(\rho_m + \rho_\Lambda) - \frac{kc^2}{R^2a^2}$$

and the right-hand side vanishes if we take a positively curved universe, $k = +1$, with radius

$$Ra = \frac{c^4}{8\pi G\rho_\Lambda} = \frac{c^2}{\Lambda} \tag{1.63}$$

This is the *Einstein static universe*. It is unstable. If a is a little smaller than the critical value (1.63) then $\rho_m \sim a^{-3}$ is a little larger and the acceleration equation (1.62) says that a will decrease further. Similarly, if a is larger than the critical value it will increase further.

1.3.4 How We Found Our Place in the Universe

In 1543, Copernicus argued that we do not sit at the centre of the universe. It took many centuries for us to understand where we do, in fact, sit.

Thomas Wright was perhaps the first to appreciate the true vastness of space. In 1750, he published "An original theory or new hypothesis of the universe", suggesting that the Milky Way, the band of stars that stretches across the sky, is in fact a "flat layer of stars" in which we are embedded, looking out. He further suggested that cloudy spots in the night sky, known as nebulae, are other galaxies, "too remote for even our telescopes to reach".

Wright was driven by poetry and art as much as astronomy and science and his book is illustrated by glorious pictures. His flights of fantasy led him to guesstimate that there are $3,888,000$ stars in the Milky Way, and 60 million planets. We now know, of course, that Wright's imagination did not stretch far enough: he underestimated the number of stars in our galaxy by 7 orders of magnitude.

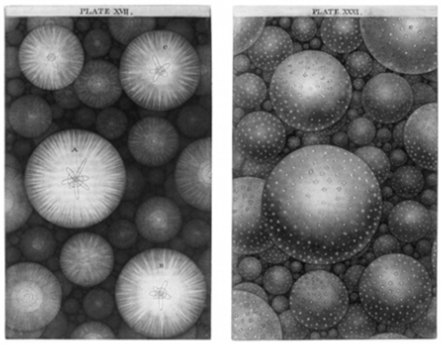

Figure 16: The wonderful imagination of Thomas Wright

Wright's suggestion that spiral nebulae are far flung galaxies, similar to our own Milky Way, was not met with widespread agreement. As late as 1920, many astronomers held that these nebulae were part of the Milky Way itself. Their argument was simple: if these were individual galaxies, or "island universes" as Kant referred to them, then they would lie at distances too vast to be credible.

The dawning realisation that our universe does indeed spread over such mind boggling distances came only with the discovery of redshifts. The American astronomer Vesto Slipher was the first to measure redshifts in 1912. He found spiral nebulae with both blueshifts and redshifts, some moving at speeds which are much too fast to be gravitationally bound to the Milky Way. Yet Slipher did not appreciate the full significance of his observations.

A number of other astronomers improved on Slipher's result, but the lion's share of the credit ended up falling into the lap of Edwin Hubble. His data, first shown in 1925, convinced everyone that the nebulae do indeed lie far outside our galaxy at distances of hundreds of kiloparsecs. Subsequently, in 1929 he revealed further data and laid claim to the law $\mathbf{v} = H\mathbf{x}$ that bears his name. For this, he is often said to have discovered the expanding universe. Yet strangely Hubble refused to accept this interpretation of his data, claiming as late as 1936 that "expanding models are definitely inconsistent with the observations that have been made".

It fell to theorists to put the pieces together. A framework in which to discuss the entire cosmos came only with the development of general relativity in 1915. Einstein himself was the first to apply relativity to the universe as a whole. In 1917, driven by a philosophical urge for an unchanging universe, he introduced the cosmological constant to apply a repulsive pressure which would counteract the gravitational attraction of matter, resulting in the static spacetime that we met in (1.63). After Einstein's death, the physicist Gammow gave birth to the famous "biggest blunder" legend, stating

> "Einstein remarked to me many years ago that the cosmic repulsion idea was the biggest blunder he had made in his entire life."

Many other physicists soon followed Einstein. First out of the blocks was the dutch astronomer Willem de Sitter who, in 1917, published the solution that now bears his name, describing a spacetime with positive cosmological constant and no matter. de Sitter originally wrote the solution in strange coordinates, which made him think that his spacetime was static rather than expanding. He was then surprised to discover that signals between distant observers are redshifted. Both Slipher and Hubble referred to their redshift observations as the "de Sitter effect".

In St Petersburg, an applied mathematician-cum-meteorologist called Alexander Friedmann was also looking for solutions to the equations of general relativity. He derived his eponymous equation in 1922 and found a number of solutions, including universes which contracted and others which expanded indefinitely. Remarkably, at the end of his paper he pulls an estimate for the energy density of the universe out

of thin air, gets it more or less right, and comes up with an age of the universe of 10 billion years. Sadly his work was quickly forgotten and three years later Friedmann died. From eating a pear. (No, really.)

The first person to understand the big picture was a Belgian, Catholic priest called Georges Lemaître. In 1927 he independently reproduced much of Friedmann's work, finding a number of further solutions. He derived Hubble's law (two years before Hubble's observations), extracting the first derivation of H_0 in the process and was, moreover, the first to connect the redshifts predicted by an expanding universe with those observed by Slipher and Hubble. For this reason, many books refer to the FRW metric as the FLRW metric. Although clearly aware of the significance of his discoveries, he chose to publish them in French in "Annales de la Société Scientifique de Bruxelles", a journal which was rather far down the reading list of most physicists. His work only became publicised in 1931 when a translation was published in the Monthly Notices of the Royal Astronomical Society, by which time much of the credit had been bagged by Hubble. Lemaître, however, was not done. Later that same year he proposed what he called the "hypothesis of the primeval atom", these days better known as the Big Bang theory. He was also the first to realise that the cosmological constant should be identified with vacuum energy.

We have not yet met R and W. The first is Howard Robertson who, in 1929, described the three homogeneous and isotropic spaces. This work was extended in 1935 by Roberston and, independently, by Arthur Walker, who proved these are the only possibilities.

Despite all of these developments, there was one particularly simple solution that had fallen through the cracks. It fell to Einstein and de Sitter to fill this gap. In 1932, when both were visitors at Caltech, they collaborated on a short, 2 page paper in which they described an expanding FRW universe with only matter. The result is the Einstein-de Sitter universe that we met in (1.49). Apparently neither thought very highly of the paper. Eddington reported a conversation with Einstein, who shrugged off this result with

> "I did not think the paper very important myself, but de Sitter was keen on it."

On hearing this, de Sitter wrote to Eddington to put the record straight,

> "You will have seen the paper by Einstein and myself. I do not myself consider the result of much importance, but Einstein seemed to think it was."

This short, unimportant paper, unloved by both authors, set the basic framework for cosmology for the next 60 years, until the cosmological constant was discovered in the late 1990s. As we will see in the next section, it provides an accurate description of the expansion of the universe for around 10 billion years of its history.

1.4 Our Universe

The time has now come to address the energy content and geometry of our own universe. We have come across a number of different entities that can contribute to the energy density of a universe. The three that we will need are

- Conventional matter, with $\rho_m \sim a^{-3}$
- Radiation, with $\rho_r \sim a^{-4}$
- A cosmological constant, with ρ_Λ constant.

We will see that these appear in our universe in somewhat surprising proportions.

Critical Density

Recall from Section 1.3.2 that in a flat universe the total energy density today must sum to match the Hubble constant. This is referred to as the *critical energy density*,

$$\rho_{\rm crit,0} = \frac{3c^2}{8\pi G} H_0^2 \tag{1.64}$$

We use this to define dimensionless *density parameters* for each fluid component,

$$\Omega_w = \frac{\rho_{w,0}}{\rho_{\rm crit,0}}$$

We have not included a subscript 0 on the density parameters but, as the definition shows, they refer to the fraction of energy observed today. Cosmologists usually specify the energy density in our Universe in terms of these dimensionless numbers Ω_w.

By design, the dimensionless density parameters sum to

$$\sum_{w=m,r,\Lambda} \Omega_w = 1 + \frac{kc^2}{R^2 H_0^2}$$

In particular, if we are to live in a flat universe then we must have $\sum_w \Omega_w = 1$. Any excess energy density, with $\sum_w \Omega_w > 1$ means that we necessarily live in a positively curved universe with $k = +1$. Any deficit in the energy, with $\sum_w \Omega_w < 1$ gives rise to a negatively curved, $k = -1$ universe.

It is sometimes useful to place the curvature term on a similar footing to the other energy densities. We define the energy density in curvature to be

$$\rho_k = -\frac{3kc^4}{8\pi G R^2 a^2}$$

and the corresponding density parameter as

$$\Omega_k = \frac{\rho_{k,0}}{\rho_{\rm crit,0}} = -\frac{kc^2}{R^2 H_0^2} \tag{1.65}$$

With these definitions, together with the scaling $\rho_w = \rho_{w,0}\, a^{-3(1+w)}$, the Friedmann equation

$$H^2 = \frac{8\pi G}{3c^2} \sum_{w=m,r,\Lambda} \rho_w - \frac{kc^2}{R^2 a^2}$$

can be rewritten in terms of the density parameters as

$$\left(\frac{H}{H_0}\right)^2 = \frac{\Omega_r}{a^4} + \frac{\Omega_m}{a^3} + \frac{\Omega_k}{a^2} + \Omega_\Lambda \tag{1.66}$$

One of the tasks of observational cosmology is to measure the various parameters in this equation.

1.4.1 The Energy Budget Today

After many decades of work, we have been able to measure the energy content of our universe fairly accurately. The two dominant components are

$$\Omega_\Lambda = 0.69 \quad \text{and} \quad \Omega_m = 0.31 \tag{1.67}$$

The cosmological constant, which we now know comprises almost 70% of the energy of our universe, was discovered in 1998. There are now two independent pieces of evidence. The first comes from direct measurement of Type Ia supernovae at large redshifts. (We saw the importance of supernovae in Section 1.1.5.) Similar data from 2003 is shown in Figure 17[4]. The 2011 Nobel prize was awarded to Perlmutter, Schmidt and Riess for this discovery.

The second piece of evidence is slightly more indirect, although arguably cleaner. The fluctuations in the cosmic microwave background (CMB) contain a wealth of information about the early universe. In combination with information from the distribution of galaxies in the universe, this provides separate confirmation of the results (1.67), as shown in Figure 18. (The label BAO in this figure refers "baryon acoustic oscillations"; we will briefly discuss these in Section 3.2.3.)

[4]This data is taken from R. Knopp et al., "New Constraints on Ω_m, Ω_Λ, and w from an Independent Set of Eleven High-Redshift Supernovae Observed with HST", Astrophys.J.598:102 (2003).

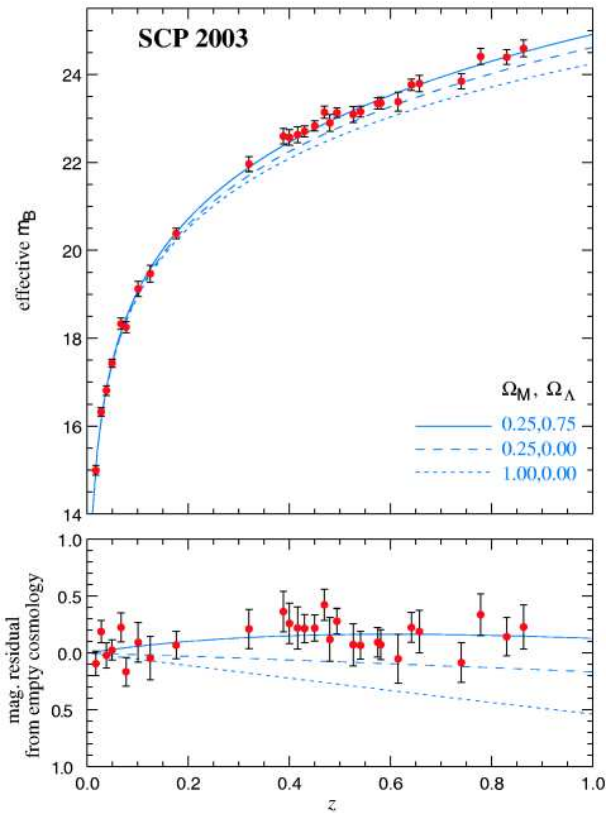

Figure 17: The redshift of a number of supernovae plotted against measured brightness. Various theoretical curves are shown for comparison.

All other contributions to the current energy budget are orders of magnitude smaller. For example, the amount of energy in photons (denoted as γ) is

$$\Omega_\gamma \approx 5 \times 10^{-5} \tag{1.68}$$

Moreover, as the universe expanded and particles lost energy and slowed, they can transition from relativistic speeds, where they count as "radiation", to speeds much less than c where they count as "matter". This happened fairly recently to neutrinos, which contribute $\Omega_\nu \approx 3.4 \times 10^{-5}$.

Finally, there is no evidence for any curvature in our universe. The bound is

$$|\Omega_k| < 0.01$$

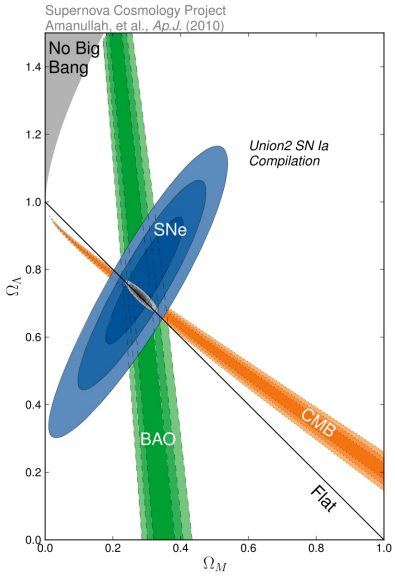

Figure 18: CMB, BAO and Supernovae results combined.

This collection of numbers, Ω_m, Ω_Λ, Ω_r and Ω_k sometimes goes by the name of the ΛCDM model, with Λ denoting the cosmological constant and CDM denoting *cold dark matter*, a subject we'll discuss more in Section 1.4.3.

The lack of any suggestion of curvature strongly suggests that we are living in a universe with $k = 0$. Given that the curvature of the universe is a dynamical variable and, as we have seen in Section 1.3.2, the choice of a flat universe is unstable, this is rather shocking. We will offer a putative explanation for the observed flatness in Section 1.5.

Energy and Time Scales

To convert the dimensionless ratios above into physical energy densities and time scales, we need an accurate measurement of the Hubble constant. Here there is some minor controversy. A direct measurement from Type IA supernovae gives[5]

$$H_0 = 74.0 \ \ (\pm 1.4) \ \mathrm{km\,s^{-1}Mpc^{-1}}$$

[5]The latest supernova data can be found in Riess et al., arXiv:1903.07603. Meanwhile, the final Planck results, extracting cosmological parameters from the CMB, can be found at arXiv:1807.06209. A different method of calibrating supernovae distances has recently found the result $H_0 = 69.8(\pm 1.7)\ \mathrm{km\,s^{-1}Mpc^{-1}}$, in much closer agreement with the CMB data; see arXiv:1907.05922.

Meanwhile, analysis of the cosmic microwave background measured by the Planck satellite puts the value at

$$H_0 = 67.4 \ \ (\pm 0.5) \ \mathrm{km\,s^{-1}Mpc^{-1}}$$

The error bars suggest a 3σ discrepancy between the two measurements. Most of the community suspect that there is some systematic issue in one of the measurements, possibly in our understanding of cepheid luminosity which is used as a calibration for the supernovae results. However, it remains a possibility that there is something important and fundamental hiding in this mismatch. Here we use $H_0 \approx 70 \ \mathrm{km\,s^{-1}Mpc^{-1}} = 2.3 \times 10^{-18} \ \mathrm{s^{-1}}$.

From a knowledge of the Hubble constant, and with $k = 0$, the expression (1.64) tells us that the total energy density of the universe is equal to the critical density,

$$\rho_{\mathrm{crit},0} = \frac{3c^2H_0^2}{8\pi G} = 8.5 \ \times 10^{-10} \ \ \mathrm{kg\,m^{-1}s^{-2}} \tag{1.69}$$

The corresponding mass density is

$$\frac{\rho_{\mathrm{crit},0}}{c^2} = \frac{3H_0^2}{8\pi G} \approx 10^{-26} \ \mathrm{kg}\, m^{-3} \tag{1.70}$$

This is about one galaxy per cubic Mpc. Or, in more down to earth terms, one hydrogen atom per cubic metre. (Or, if you like, 10^{-68} galaxies per cubic metre!) The actual matter in the universe is, of course, fractionally less at $\rho_{m,0} = \Omega_m\rho_{\mathrm{crit},0}$.

With the universe dominated by ρ_Λ and ρ_m, the solution (1.60), given by

$$a(t) = \left(\frac{\rho_0}{\rho_\Lambda}\right)^{1/3} \sinh^{2/3}\left(\frac{\sqrt{3\Lambda}t}{2}\right)$$

offers a good description of the expansion for much of this history. Recall that, in such a solution, the Big Bang takes place at $t = 0$ while the present day is defined by

$$\sinh^2\left(\frac{\sqrt{3\Lambda}t_0}{2}\right) = \frac{\rho_\Lambda}{\rho_0}$$

Inverting this gives the age

$$t_0 = \frac{c}{\sqrt{6\pi G\rho_\Lambda}} \sinh^{-1}\left(\sqrt{\frac{\rho_\Lambda}{\rho_0}}\right) = \frac{2}{3\sqrt{\Omega_\Lambda}H_0} \sinh^{-1}\left(\sqrt{\frac{\Omega_\Lambda}{\Omega_0}}\right)$$

The various factors almost cancel out, leaving us with an age which is very close to the naive estimate (1.23)

$$t_0 \approx 0.96 \times \frac{1}{H_0} \approx 1.4 \times 10^{10} \text{ years}$$

We can also calculate the age at which the vacuum energy was equal to the energy in matter (1.61). We get

$$t_{\rm eq} = \frac{2}{3\sqrt{\Omega_\Lambda} H_0} \sinh^{-1}(1) \approx 0.7 \times \frac{1}{H_0} \approx 0.98 \times 10^{10} \text{ years}$$

or about 4 billion years ago. To put this in perspective, the Earth is around 4.5 billion years old, and life started to evolve (at least) 3.5 billion years ago. In the grand scheme of things, equality between matter energy density and the cosmological constant occurred very recently.

Throughout these lectures, we will often use redshift z, rather than years, to refer to the time at which some event happened. Recall the the redshift is defined as (1.21)

$$1 + z = \frac{1}{a}$$

This means that at redshift z,the universe was $1/(1+z)^{\rm th}$ its present size. This has the advantage that it's very easy to compute certain numbers in terms of z. For example, the equality of the cosmological constant and matter occurred when $\rho_m = \rho_\Lambda$ which, in terms of todays fractional energy density ,means that $\Omega_m/a^3 = \Omega_\Lambda$. Plugging in the numbers gives $z = 0.3$.

Matter-Radiation Equality

Today, radiation is an almost negligible part of the total energy density. However, this wasn't always the case. Because $\rho_r \sim 1/a^4$, as we go backwards in time the energy density in radiation grows much faster than matter, with $\rho \sim 1/a^3$, or the cosmological constant. We can ask: when do we have matter-radiation equality? In terms of redshift this requires

$$\frac{\Omega_m}{a^3} = \frac{\Omega_r}{a^4}$$

Here there is a small subtlety because neutrinos transition from relativistic to non-relativistic during this period. If we include the present day neutrino density as radiation, then we have $\Omega_r \approx 8.4 \times 10^{-5}$, which gives matter-radiation equality at $z \approx 3700$. A more accurate assessment gives

$$z_{\rm eq} \approx 3400 \tag{1.71}$$

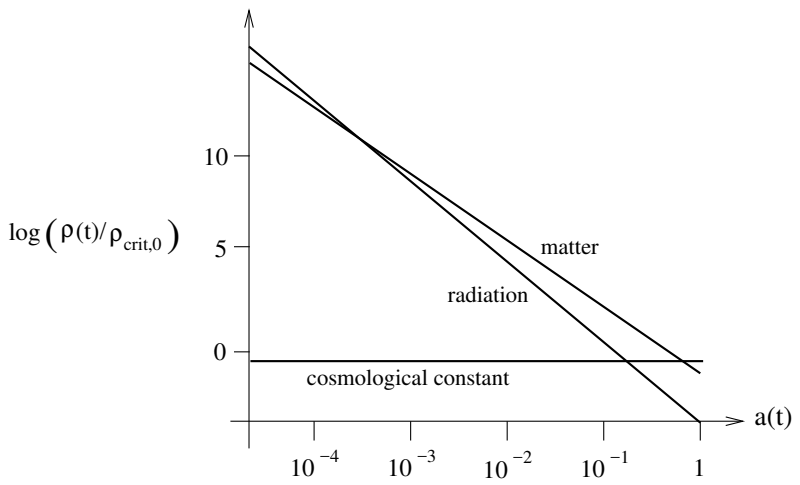

Figure 19: The evolution of the energy densities in our universe.

We can translate this into years. The universe was matter dominated for most of the time since $z = 3400$, with the cosmological constant becoming important only (relatively) recently. If we work with $a(t) = (t/t_0)^{2/3}$ as befits a matter-dominated universe, then we can trace back the evolution from the present day to get a rough estimate for the time of matter-radiation equality to be

$$t_{\rm eq} = \frac{t_0}{(1+z_{\rm eq})^{3/2}} \approx 70,000 \text{ years}$$

A more accurate calculation gives

$$t_{\rm eq} \approx 50,000 \text{ years}$$

Prior to this, the universe was radiation dominated.

A plot of the evolution of the three kinds of energy is shown in Figure 19.

1.4.2 Dark Energy

For a number of observational cosmologists, who had long been wrestling with the difficulty of reconciling the early age (1.50) of a matter dominated universe with the lifetime of stars, the discovery of the cosmological constant came as a welcome relief. However, for the more theoretical minded physicists, it was something of a bombshell.

In the comfortable world of classical physics, there is no mystery to the cosmological constant. It is, as we have seen, simply the constant energy term which we previously

neglected. However, our fundamental theories of physics are quantum. And here there is a problem, because they provide a way to estimate the size of the cosmological constant Λ.

To tell this story, we first need to adapt our units. Taking the critical energy density to be (1.69), the observed vacuum energy density is $\rho_\Lambda \approx 6 \times 10^{-10}$ J m^{-3}. However, a more natural unit of energy is not the joule, but the *electron volt*, with 1 J $\approx 6.2 \times 10^{18}$ eV. In these units,

$$\rho_\Lambda = 3.7 \times 10^9 \text{ eV m}^{-3}$$

Perhaps more surprisingly, our preferred unit of inverse length is also the electron volt! To convert from one to the other, we use the fundamental constants of nature, $\hbar c \approx 2.0 \times 10^{-7}$ eV m. Putting this together, gives

$$\hbar^3 c^3 \rho_\Lambda \approx (10^{-3} \text{ eV})^4$$

Usually, this is written in natural units, with $\hbar = c = 1$, so that

$$\rho_\Lambda \approx (10^{-3} \text{ eV})^4$$

What are our expectations for the vacuum energy? Our fundamental laws of physics are written in framework called *quantum field theory*. All quantum field theories have a term, analogous to the $+\frac{1}{2}\hbar\omega$ ground state energy of the harmonic oscillator, which contributes to the vacuum energy of the universe. However, in contrast to the harmonic oscillator, in quantum field theory the ground state energy gets contributions from all possible frequencies. Taken at face value, this integral over frequencies would appear to diverge.

To make sense of this divergence, we need to embrace a little humility. Our theories have not been tested to arbitrarily high energy scales, and surely break down at some point. The best we can say at present is that the theories make sense up to the scales tested at the LHC, which operates at energies

$$M_{\rm LHC} \sim 1 \text{ TeV} = 10^{12} \text{ eV}$$

With this conservative estimate for the validity of our theories, the most "natural" value for the vacuum energy arising from quantum field theory

$$\rho_{\rm QFT} = (10^{12} \text{ eV})^4 = 10^{60} \rho_\Lambda$$

This is not particularly close to the observed value. It is, moreover, a ridiculous number that makes no sense in the cosmological context. Such a universe would not be

conducive to forming nuclei or atoms, let alone galaxies and life. The huge discrepancy between the expected value of ρ_{QFT} and the observed value of ρ_Λ is known as the *cosmological constant problem*.

Physicists with masochistic tendencies will try to make the situation look even worse. There is some minimal, circumstantial evidence that the framework of quantum field theory holds up to the Planck scale $M_{pl} = \sqrt{\hbar c/8\pi G}$ which corresponds to the energy $M_{\rm pl}c^2 \approx 10^{19}$ GeV. In this case, we would get $\rho_{QFT} = 10^{122}\rho_\Lambda$. I'm not sure this way of stating things is particularly helpful.

The value of ρ_{QFT} is not a precise prediction of quantum field theory, but rather a ballpark figure for the natural energy scale of the theory. We are always free to just add a further arbitrary constant to the energy of the theory. In that case, there are two contributions

$$\rho_\Lambda = \rho_{QFT} + \rho_{\rm constant}$$

Apparently, the two contributions on the right must add up to give the observed value of ρ_Λ. We call this *fine-tuning*. As presented above, it looks fairly absurd: two numbers of order 10^{60} (or higher) have to coincide in the first 60 digits, but differ in the $61^{\rm st}$, leaving behind a number of order 1.

It is quite possible that there is some missing principle that we've failed to grasp that makes fine tuning less silly than it first appears. The task of finding such a mechanism is made considerably harder when we realise that there have been a number of times in the history of the universe when ρ_{QFT} abruptly changed while, presumably, $\rho_{\rm constant}$ did not. This occurs at a *phase transition*. For example, the QCD phase transition, where quarks which were once free became trapped in protons and neutrons, took place in the early universe. At this moment, there was a change $\Delta\rho_{QFT} \sim (100\text{ MeV})^4$. Still earlier, the electroweak phase transition, where the Higgs boson kicks in and gives mass to fundamental particles, should have resulted in a change of $\Delta\rho_{QFT} \sim (100\text{ GeV})^4$. In other words, any putative cancellation mechanism must conspire to give a tiny cosmological constant ρ_Λ at the end of the life of the universe, not at the beginning.

Given these difficulties, most physicists in the $20^{\rm th}$ century buried their heads in the sand and assumed that there must be some deep principle that sets the cosmological constant to zero. No such principle was found. In the $21^{\rm st}$ century, we have a much harder job. We would like a deep principle that sets the late-time cosmological constant to $\rho_\Lambda \sim (10^{-3}\text{ eV})^4$. Needless to say, we haven't found that either.

If this wasn't bad enough, there is yet another issue that we should confront. The value of the current vacuum energy is remarkably close to the energy in matter. Why? As illustrated in Figure 19, these energy densities scale very differently and we would naively expect that they differ by orders of magnitude. Why are Ω_m and Ω_Λ so very close today? This is known as the *coincidence problem*. We have no good explanation.

The A-Word

As we saw above, a naive application of quantum field theory suggests a ludicrous value for the cosmological constant, one that results in an expansion so fast that not even atoms have a chance to form from their underlying constituents. Given this, we could ask the following question: what is the maximum value of the cosmological constant that still allows complex structures to evolve? For example, what is the maximum allowed value of Λ that allows galaxies to form?

It turns out that the upper bound on Λ depends on the strength of the initial seeds from which the galaxies grew. At very early times, there are small variations $\delta\rho$ in the otherwise homogeneous universe. As we will discuss in more detail in Section 3, in our universe these seeds have size $\delta\rho/\rho \sim 10^{-5}$. Let us fix this initial condition, and then ask again: how big can the cosmological constant be?

We will present this calculation in Section 3.3.4. The answer is quite striking: the scale of the vacuum energy is pretty much the maximum it could be. If ρ_Λ were bigger by an order of magnitude or so, then no galaxies would form, presumably making it rather more difficult for life to find a comfortable foothold in the universe.

What to make of this observation? One possibility is to shrug and move on. Another is to weave an elaborate story. Suppose that our observable universe is part of a much larger structure, a "multiverse" in which different domains exhibit different values of the fundamental parameters, or perhaps even different laws of physics. In this way, the cosmological constant is not a fundamental parameter which we may hope to predict, but rather an environmental parameter, no different from, say, the distance between the Earth and the Sun. We should not be shocked by its seemingly small value because, were it any higher, we wouldn't be around to comment on it. Such reasoning goes by the name of the *anthropic principle*.

The anthropic explanation for the cosmological constant may be correct. But, in the absence of any testable predictions, discussions of this idea rapidly descend into a haze of sophomoric tedium. Trust me: there are better things to do with your life. (Like find a proper explanation.)

A Rebranding: Dark Energy

Given our manifest befuddlement about all things Λ, it is prudent to wonder if ρ_Λ is actually a cosmological constant at all. Perhaps it is some other form of fluid, with an equation of state $w \approx -1$, rather than precisely $w = -1$. More interesting, it may be a fluid whose equation of state evolves over time. (We will meet behaviour like this in Section 1.5.) I stress that there are no compelling theoretical reasons to believe that this is the case, and nor does it alleviate the need to explain why ρ_{QFT} does not gravitate. Nonetheless, this is clearly an area where we are totally at sea and we should be open to such possibilities. For these reasons, the mysterious 70% of the energy in the universe is often referred to as *dark energy*.

1.4.3 Dark Matter

Our embarrassing ignorance of the universe we call home is further illustrated if we delve a little deeper into the $\Omega_m = 0.31$ energy in matter. Of this, the amount that we understand is

$$\Omega_B \approx 0.05 \tag{1.72}$$

This is the energy in matter made from atoms in the periodic table. The B in Ω_B stands for "baryons", which are protons and neutrons. This is appropriate because the mass in electrons is negligible in comparison.

The remaining matter energy is in the form of *cold dark matter*,

$$\Omega_{CDM} \approx 0.26$$

This is stuff that we have not (yet?) created here on Earth. The "cold" refers to the fact that it is non-relativistic today and, moreover, has been so for some time.

We know very little about this dark matter. We do not know if it is a single species of particle, or many. We do not know if it consists of several decoupled sectors, or just one. Given the wonderful complexity that lurks in Ω_B, it seems reasonable to assume that there is still rather a lot to learn about Ω_{CDM}.

Here we simply describe some of the evidence for the existence of dark matter. To do this, we need to construct methods to determine the mass of the large objects, such as galaxies or clusters of galaxies. These are small enough for us to ignore the expansion of the universe so, for the rest of this section, we will work in flat space.

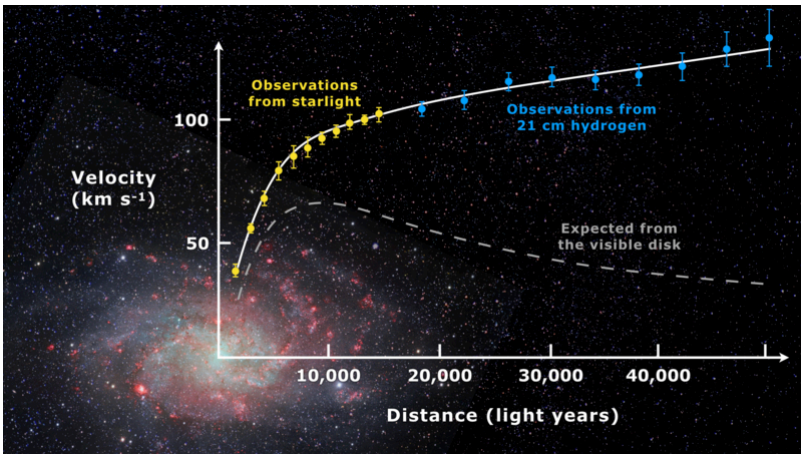

Figure 20: The rotation curve of galaxy M33. Image taken from Wikipedia.

Galaxy Rotation Curves

At the galactic scale, rotation curves provide a clean way to measure mass. This method was pioneered by Vera Rubin and her collaborator Kent Ford in the 1960s and 70s.

For a quick and dirty sketch of the idea, we will assume spherical symmetry. A quick glance at a typical spiral galaxy shows this is a poor approximation, at least for the visible matter, but it will suffice to get the basic idea across. The centrifugal acceleration of a star, orbiting at distance r from the galactic centre, must be provided by the gravitational force,

$$\frac{v^2}{r} = \frac{GM(r)}{r^2}$$

where $M(r)$ is the mass enclosed inside a sphere of radius r. We learn that we expect the rotational speed to vary as

$$v(r) = \sqrt{\frac{GM(r)}{r}}$$

Far from the bulk of the galaxy, we would expect that $M(r)$ is constant, so the velocity drops off as $v \sim \sqrt{1/r}$. This is not what is observed. The rotation speeds can be measured from the edge of the galaxy by studying interstellar gas, in particular the 21cm line of hydrogen. (The origin of this line was discussed in the Atomic Physics section of the Lectures on Topics in Quantum Mechanics..) One finds that the rotation remains more or less constant very far from what appears to be the edge of the galaxy. This suggests that the mass continues to grow as $M(r) \sim r$ far from the observable galaxy. This is known as the dark matter halo.

The Virial Theorem and Galaxy Clusters

The *virial theorem* offers a clever method of weighing a collection of objects that are far away.

Virial Theorem: A collection of N particles, with masses m_i and positions $\mathbf{x}_i$, interact through a gravitational potential

$$V = \sum_{i<j} V_{ij} = \sum_{i<j} -\frac{Gm_i m_j}{|\mathbf{x}_i - \mathbf{x}_j|} \tag{1.73}$$

We will assume that the system is gravitationally bound, and that the positions $\mathbf{x}_i$ and velocities $\dot{\mathbf{x}}_i$ are bounded for all time. We will further assume that the time average of the kinetic energy T and potential energy V are well defined. Then

$$\overline{T} = -\frac{1}{2}\overline{V}$$

where the bar denotes time average (a quantity we will define more precisely below).

Proof: We start by defining something akin to the moment of inertia,

$$I = \frac{1}{2}\sum_i m_i \mathbf{x}_i \cdot \mathbf{x}_i \quad \Rightarrow \quad \dot{I} = \sum_i \mathbf{p}_i \cdot \mathbf{x}_i \tag{1.74}$$

with $\mathbf{p}_i$ the momentum of the i^{th} particle. The quantity $\dot{I}$ is known as the *virial*. Note that, in contrast to the potential V, both I and $\dot{I}$ depend on our choice of origin. The correct choice is to pick this origin to be the centre of mass. The time derivative of the virial is

$$\ddot{I} = \sum_i \dot{\mathbf{p}}_i \cdot \mathbf{x}_i + \sum_i \mathbf{p}_i \cdot \dot{\mathbf{x}}_i = \sum_i \mathbf{F}_i \cdot \mathbf{x}_i + 2T$$

where, in the second equality, we have used the definition of kinetic energy T and Newton's force law $\mathbf{F}_i = \dot{\mathbf{p}}_i$. The force $\mathbf{F}_i$ on the i^{th} particle is determined by the potential V_{ij} by

$$\begin{aligned}
\mathbf{F}_i = -\sum_{j\neq i} \nabla_i V_{ij} \quad \Rightarrow \quad \sum_i \mathbf{F}_i \cdot \mathbf{x}_i &= -\sum_{i<j} \nabla_i V_{ij} \cdot \mathbf{x}_i - \sum_{j<i} \nabla_i V_{ij} \cdot \mathbf{x}_i \\
&= -\sum_{i<j} \nabla_i V_{ij} \cdot \mathbf{x}_i - \sum_{i<j} \nabla_j V_{ji} \cdot \mathbf{x}_j \\
&= -\sum_{i<j} \nabla_i V_{ij} \cdot (\mathbf{x}_i - \mathbf{x}_j)
\end{aligned}$$

where, in the second step, we simply swapped the dummy indices i and j and, in the third step, we used $V_{ij} = V_{ji}$ and $\nabla_i V_{ij} = -\nabla_j V_{ij}$. But now we can use the explicit form of the potential (1.73) to find

$$-\sum_{i<j} \nabla_i V_{ij} \cdot (\mathbf{x}_i - \mathbf{x}_j) = \sum_{i<j} V_{ij} = V$$

We learn that the time variation of the virial is

$$\ddot{I} = V + 2T$$

At this point we take the time average, defined by

$$\overline{X} = \lim_{t\to\infty} \frac{1}{t} \int_0^t X(t')\, dt'$$

The time average of all these quantities is assumed to be well-defined. But,

$$\overline{\frac{d\dot{I}}{dt}} = \lim_{t\to\infty} \frac{\dot{I}(t) - \dot{I}(0)}{t} = 0$$

Note that the last step follows only if the virial (1.74) is measured relative to the centre of mass, otherwise the positions $\mathbf{x}_i$ will have a drift linear in time. We're left with the promised virial theorem $\overline{V} + 2\overline{T} = 0$. □

As an aside: the virial theorem also holds in other contexts. For example, a proof using the variational method can be found in the Lectures on Topics in Quantum Mechanics.

The virial theorem can be used to estimate the mass of any collection of objects that satisfy the assumptions of the theorem. Roughly speaking, this holds when the objects have reached something akin to thermodynamic equilibrium. In 1933, Zwicky used this technique to estimate the mass of the Coma cluster, shown in the figure, a conglomerate of a few thousand galaxies.

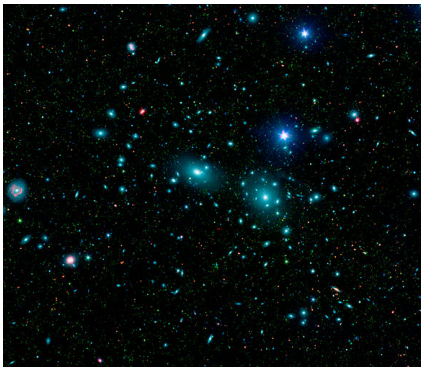

Figure 21: Coma cluster.

We will make a few simplifying assumptions. First we will assume that there are N galaxies, all of the same mass m. (We can do better, but this will serve our purposes.) Second, we will assume that the system is "self-averaging", which

means that the average over many galaxies is a proxy for averaging over time so that, for example,

$$\bar{T} \approx \langle T \rangle = \frac{1}{2N} \sum_{i=1}^{N} m v_i^2$$

This has the advantage that we don't need to wait several billion years to perform the time average. The virial theorem then reads

$$2\langle T \rangle = m\langle v^2 \rangle \approx \langle V \rangle \approx \frac{1}{2} G m^2 N \left\langle \frac{1}{r} \right\rangle$$

where $\langle 1/r \rangle$ is the average inverse distance between galaxies and, in the last step, we have replaced $N - 1$ with N. This then gives an expression for the total mass of the galaxy cluster,

$$Nm \approx \frac{2\langle v^2 \rangle}{G\langle 1/r \rangle} \tag{1.75}$$

The right-hand-side contains quantities that we can measure, giving us an estimate for the mass of the cluster. (Strictly speaking, we can measure v_{redshift}, the velocity in the line of sight. If we further assume spherical symmetry, we have $\langle v^2 \rangle = 3\langle v^2_{\text{redshift}} \rangle$.)

There is a much simpler way to compute the mass in each galaxy: simply count the number of stars. In practice, this is done by measuring the luminosity. This provides two very different ways to determine the mass and we can compare the two. One typically finds that the virial mass is greater than the luminosity mass by a factor of a couple of hundred. The difference is made up by what Zwicky referred to as *Dunkle Materie*, or dark matter.

Other Evidence

There are a number of other pieces of evidence, all of which consistently point to the existence of dark matter. The mathematics underlying these requires more than just Newtonian dynamics so, for now, we will replace the maths with some pretty pictures.

- Gravitational Lensing: A classic prediction of general relativity is that light bends as it passes heavy objects. Furthermore, the image gets distorted, a phenomenon known as *gravitational lensing*. Sometimes this happens in a spectacular fashion, as shown in the picture on the left, where the image of a background galaxy is distorted into the blue arcs by the cluster in the foreground. Even small distortions of this kind allow us accurately determine the mass of the cluster in

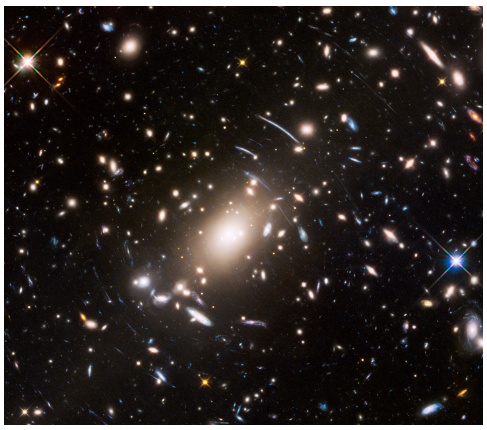

Figure 22: Abell S1063 cluster.

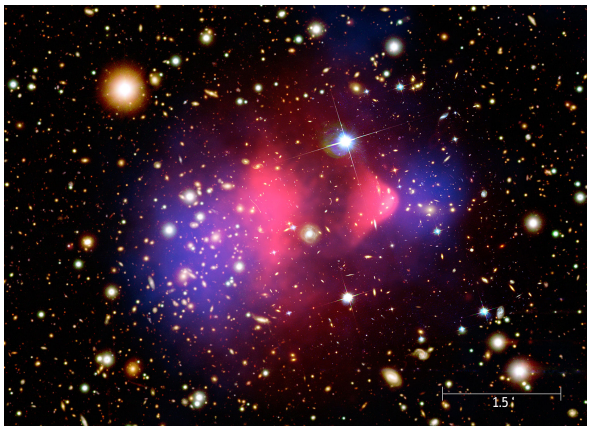

Figure 23: The bullet cluster.

the foreground. You will not be surprised to hear that the mass greatly exceeds that seen in visible matter.

The bullet cluster, shown in the right-hand figure, provides a particularly dramatic example of gravitational lensing. This picture shows two sub-clusters of galaxies which are thought to have previously collided. There are three types of matter shown in the picture: stars which you can see, hot gas which is observed in x-rays and is shown in pink, and the distribution of mass detected through gravitational lensing shown in blue. The stars sit cleanly in two distinct sub-clusters because individual galaxies have little chance of collision. In contrast, most of the baryonic matter sits in clouds of hot gas which interact fairly strongly as the clusters collide, slowing the gas and leaving it displaced from the stars as shown in the figure. But most of the matter, as detected through gravitational lensing, is dark and this, like the galaxies, has glided past each other seemingly unaffected by the collision. The interpretation is that dark matter interacts weakly, both with itself and with baryonic matter.

- BBN: The observations described above show clearly that on the scale of both galaxies and clusters of galaxies there is more matter than can be detected by electromagnetic radiation. This alone is not sufficient to tell us that dark matter must be composed of some new unknown particle. For example, it could be in the form of failed stars ("jupiters"). There is, however, compelling evidence that this is not the case, and dark matter is something more exotic.

  The primary evidence comes from *Big Bang nucleosynthesis* (BBN), an impressively accurate theory of how the light elements were forged in the early universe. It turns out that the relative abundance of different elements depends on the

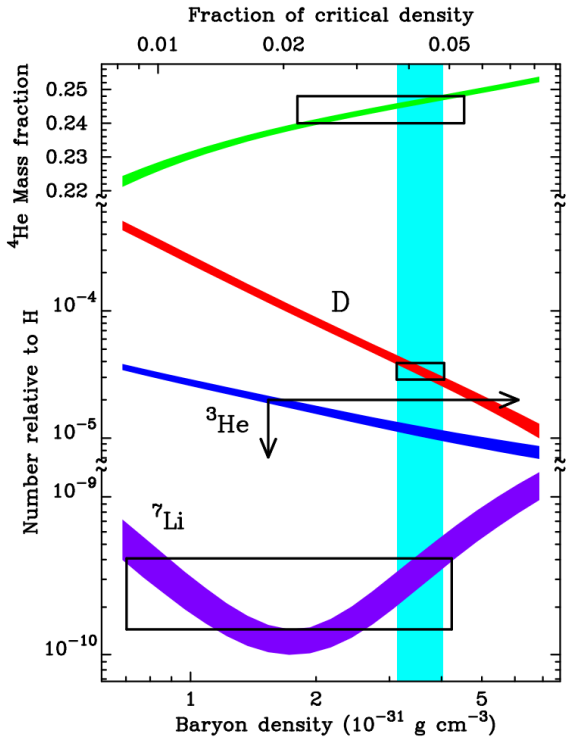

Figure 24: The relative abundance of light elements forged in the early universe, as a function of the overall baryon density.

total amount of baryon matter. In particular, the relative abundance of deuterium changes rapidly with baryon density. This is shown[6] in Figure 24, with the horizontal turquiose bar fixed by observations of relative abundance. (The black boxes show the relative abundance of each element, with error bars, and the corresponding constraint on the baryon density.) This tells us that the total amount of baryonic matter is just a few percent of the total energy density. We will describe some aspects of BBN in Section 2.5.3.

- Structure formation: The CMB tells us that the very early universe was close to homogeneous and isotropic, with fluctuations in the energy of the order of $\delta\rho/\rho \sim 10^{-5}$. Yet today, these tiny fluctuations have grown into the clusters, galaxies and stars that we see around us. How did this happen?

  It turns out that there this can not be achieved by baryonic matter alone. In the

[6]This figure is taken from Burles, Nollett and Turner, *Big-Bang Nucleosynthesis: Linking Inner Space and Outer Space*", astro-ph/99033.

fireball of the Big Bang, baryonic matter is coupled to photons and these provide a pressure which suppresses gravitational collapse. This collapse can only proceed after the fireball cools and photons decouple, an event which takes place around 300,000 years after the Big Bang. This does not leave enough time to form the universe we have today. Dark matter, however, has no such constraints. It decouples from the photons much earlier, and so its density perturbations can start to grow, forming gravitational wells into which visible matter can subsequently fall. We will tell this story in Section 3.

- CMB: As we mentioned above, baryonic matter and dark matter behave differently in the early universe. Dark matter is free to undergo gravitational collapse, while baryonic matter is prevented from doing so by the pressure of the photons. These differences leave their mark on the fireball, and this shows up in the fluctuations etched in the microwave background. This too will be briefly described in Section 3.

1.5 Inflation

We have learned that our universe is a strange and unusual place. The cosmological story that emerged above has a number of issues that we would like to address. Some of these – most notably those related to dark matter and dark energy – have yet to be understood. But there are two puzzles that do have a compelling solution, known as *cosmological inflation*. The purpose of this section is to first describe the puzzles, and then describe the solution.

1.5.1 The Flatness and Horizon Problems

The first puzzle is one we've met before: our universe shows no sign of spatial curvature. We can't say for sure that it's exactly flat but observations bound the curvature to be $|\Omega_k| < 0.01$. A universe with no curvature is a fixed point of the dynamics, but it is an *unstable* fixed point, and any small amount of curvature present in the early universe should have grown over time. At heart, this is because the curvature term in the Friedmann equation scales as $1/a^2$ while both matter and radiation dilute much faster, as $1/a^3$ and $1/a^4$ respectively.

Let's put some numbers on this. We will care only about order of magnitudes. We ignore the cosmological constant on the grounds that it has been irrelevant for much of the universe's history. As we saw in Section 1.4.1, for most of the past 14 billion years the universe was matter dominated. In this case,

$$\frac{\rho_k(t)}{\rho_m(t)} = \frac{\rho_{k,0}}{\rho_{m,0}}\, a \quad \Rightarrow \quad \Omega_k(t) = \frac{\Omega_{k,0}}{\Omega_{m,0}} \frac{\Omega_m(t)}{1+z}$$

where, for once, we have defined time-dependent density parameters $\Omega_w(t)$ and, correspondingly, added the subscript $\Omega_{m,0}$ to specify the fractional density today. This formula holds all the way back to matter-radiation equality at $t = t_{\rm eq}$ where $\Omega_m(t_{\rm eq}) \approx 1/2$ (the other half made up by radiation) and $z \approx 3000$. Using the present day value of $\Omega_{k,0}/\Omega_{m,0} \lesssim 10^{-2}$, we must have

$$|\Omega_k(t_{\rm eq})| \leq 10^{-6}$$

At earlier times, the universe is radiation dominated. Now the relevant formula is

$$\frac{\rho_k(t)}{\rho_r(t)} = \frac{\rho_{k,\rm eq}}{\rho_{r,\rm eq}}\frac{a^2}{a_{\rm eq}^2} \quad \Rightarrow \quad \Omega_k(t) = \frac{\Omega_k(t_{\rm eq})}{\Omega_r(t_{\rm eq})}\frac{(1+z_{\rm eq})^2}{(1+z)^2}\Omega_r(t)$$

We can look, for example, at the flatness of the universe during Big Bang nucleosynthesis, a period which we understand pretty well. As we will review in Section 2, this took place at $z \approx 4 \times 10^8$. Here, the curvature must be

$$|\Omega_k(t_{\rm BBN})| \leq 10^{-16}$$

We have good reason to trust our theories even further back to the electroweak phase transition at $z \approx 10^{15}$. Here, the curvature must be

$$|\Omega_k(t_{\rm EW})| \leq 10^{-30}$$

These are small numbers. Why should the early universe be flat to such precision? This is known as the *flatness problem*.

The second puzzle is even more concerning. As we have mentioned previously, and will see in more detail in Section 2, the universe is filled with radiation known as the cosmic microwave background (CMB). This dates back to 300,000 years after the Big Bang when the universe cooled sufficiently for light to propagate.

The CMB is almost perfectly uniform and isotropic. No matter which direction we look, it has the same temperature of 2.725 K. However, according to the standard cosmology that we have developed, these different parts of the sky sat outside each others particle horizons at the time the CMB was formed. This concept is simplest to see in conformal time, as shown in Figure 25.

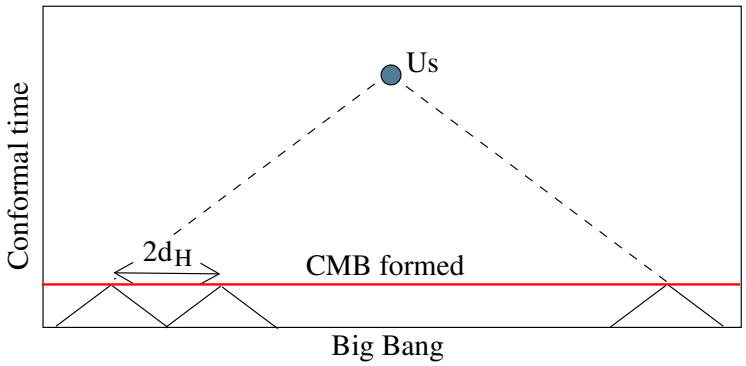

Figure 25: The horizon problem: different regions of the CMB are causally disconnected at the time it was formed.

We can put some numbers on this. For a purely matter-dominated universe, with $a(t) = (t/t_0)^{2/3}$, the particle horizon (1.24) at time t is defined by

$$d_H(t) = c\, a(t) \int_0^t \frac{dt'}{a(t')} = 3ct$$

We use $H(t) = 2/3t = H_0/a(t)^{3/2}$ to write this as

$$d_H(z) = \frac{2cH_0^{-1}}{(1+z(t))^{3/2}} \tag{1.76}$$

We will see in Section 2.3 that the CMB is formed when $z \approx 1100$. We would like to know how large the particle horizon (1.76) looks in the sky today. In the intervening time, the distance scale $d_H(z)$ has been stretched by the expansion of the universe to $(1+z)d_H(z)$. Meanwhile, this should be compared to the particle horizon today which is $d_H(t_0) = 2cH_0^{-1}$. From this, we learn that the distance $d_H(z)$ today subtends an angle on the sky given by

$$\theta \approx \frac{(1+z)d_H(z)}{d_H(t_0)} \approx \frac{1}{\sqrt{1100}} \approx 0.03 \text{ rad} \quad \Rightarrow \quad \theta \approx 1.7^\circ$$

Assuming the standard cosmology described so far, patches of the sky separated by more than $\sim 1.7^\circ$ had no causal contact at the time the CMB was formed. We would naively expect to see significant variations in temperature over the sky on this scale, but instead we see the same temperature everywhere we look. It is very hard to envisage how different parts of the universe could have reached thermal equilibrium without ever being in causal contact. This is known as the *horizon problem*.

Ultimately, the two problems above are both concerned with the initial conditions in the universe. We should be honest and admit that we're not really sure what the rules of the game are here. If you're inclined to believe in a creator, you might find it plausible that she simply stipulated that the universe was absolutely flat, with constant energy density everywhere in space at some initial time $t = \epsilon$. It's not the kind of explanation that scientists usually find compelling, but you might think it has a better chance to convince in this context.

However, there is a more nuanced version of the horizon problem which makes the issue significantly more acute, and renders the "God did it" explanation significantly less plausible. Somewhat ironically, this difficulty arises when we appreciate that the CMB is not completely uniform after all. It contains tiny, but important anisotropies. There are small fluctuations in temperature at about 1 part in 10^5. Furthermore, there are also patterns in the polarisation of the of the light in the CMB. And, importantly, the polarisation and temperature patterns are correlated. These correlations – which go by the uninspiring name of "TE correlations" – are the kind of thing that arises through simple dynamical processes in the early universe, such as photons scattering off electrons. But observations reveal that there are correlations over patches of the sky that are as large as 5°.

These detailed correlations make it more difficult to appeal to a creator without sounding like a young Earth creationist, arguing that the fossil record was planted to deceive us. Instead, the observations are clearly telling us that there were dynamical processes taking place in the early universe but, according to our standard FRW cosmology, these include dynamical processes that somehow connect points that were not in causal contact. This should make us very queasy. If we want to preserve some of our most cherished ideas in physics – such as locality and causality – it is clear that we need to do something that changes the causal structure of the early universe, giving time for different parts of space to communicate with each other.

1.5.2 A Solution: An Accelerating Phase

There is a simple and elegant solution to both these problems. We postulate that the very early universe underwent a period of accelerated expansion referred to as *inflation*. Here "very early" refers to a time before the electroweak phase transition, although we cannot currently date it more accurately than this. An accelerating phase means

$$a(t) \sim t^n \quad \text{with } n > 1 \tag{1.77}$$

Alternatively, we could have a de Sitter-type phase with $a(t) \sim e^{H_{\rm inf}t}$ with constant $H_{\rm inf}$. This is exactly the kind of accelerating phase that we are now entering due to

the cosmological constant. However, while the present dark energy is $\rho_\Lambda \sim (10^{-3}\text{ eV})^4$, the dark energy needed for inflation is substantially larger, with $\rho_{\text{inflation}} \geq (10^3\text{ GeV})^4$ and, in most models, closer to $(10^{15}\text{ GeV})^4$.

Let's see why such an inflationary phase would solve our problems. First, the horizon problem. The particle horizon is defined as (1.24),

$$d_H(t) = c\,a(t)\int_0^t \frac{dt'}{a(t')}$$

It is finite only if the integral converges. This was the case for a purely matter (or radiation) dominated universe, as we saw in (1.76). But, for $a(t) \sim t^n$ we have

$$\int_0^t \frac{dt'}{a(t')} \sim \int_0^t \frac{dt'}{t'^n} \to \infty \quad \text{if } n > 1$$

This means that an early accelerating phase buys us (conformal) time and allows far flung regions of the early universe to be in causal contact.

An inflationary phase also naturally solves the flatness problem. An inflationary phase of the form (1.77) must be driven by some background energy density that scales as

$$\rho_{\text{inf}} \sim \frac{1}{a^{2/n}}$$

which, for $n > 1$, clearly dilutes away more slowly than the curvature $\rho_k \sim 1/a^2$. This means that, with a sufficiently long period of inflation, the spatial curvature can be driven as small as we like. Although we have phrased this in terms of energy densities, there is a nice geometrical intuition that underlies this: if you take any smooth, curved manifold and enlarge it, then any small region looks increasingly flat.

This putative solution to the flatness problem also highlights the pitfalls. In the inflationary phase, the curvature ρ_k will be driven to zero but so too will the energy in matter ρ_m and radiation ρ_r. Moreover, we'll be left with a universe dominated by the inflationary energy density ρ_{inf}. To avoid this, the mechanism that drives inflation must be more dynamic than the passive fluids that we have considered so far. We need a fluid that provides an energy density ρ_{inf} for a suitably long time, allowing us to solve our problems, but then subsequently turns itself off! Or, even better, a fluid that subsequently converts its energy density into radiation. Optimistic as this may seem, we will see that there is a simple model that does indeed have this behaviour.

How Much Inflation Do We Need?

We will focus on the horizon problem. For simplicity, we will assume that the early universe undergoes an exponential expansion with $a(t) \sim e^{H_{\rm inf}t}$. Suppose that inflation lasts for some time T. If, prior to the onset of inflation, the physical horizon had size d_I then, by the end of inflation, this region of space has been blown up to $d_F = e^{H_{\rm inf}T} d_I$. We quantify the amount of inflation by $N = H_{\rm inf}T$ which we call the number of *e-folds*.

Subsequently, scales that were originally at d_I grow at a more leisurely rate as the universe expands. If the end of inflation occurred at redshift $z_{\rm inf}$, then

$$d_{\rm now} = e^N (1 + z_{\rm inf}) d_I$$

We will see that $z_{\rm inf}$ is (very!) large, and we lose nothing by writing $1 + z_{\rm inf} \approx z_{\rm inf}$. The whole point of inflation is to ensure that this length scale $d_{\rm now}$ is much larger than what we can see in the sky. This is true, provided

$$d_{\rm now} \gg c H_0^{-1} \quad \Rightarrow \quad e^N > \frac{c}{H_0 d_I} \frac{1}{z_{\rm inf}}$$

Clearly, to determine the amount of inflation we need to specify both when inflation ended, $z_{\rm inf}$, and the size of the horizon prior to inflation, d_I. We don't know either of these, so we have to make some guesses. A natural scale for the initial horizon is $d_I = c H_{\rm inf}^{-1}$, which gives

$$e^N > \frac{H_{\rm inf}}{H_0} \frac{1}{z_{\rm inf}}$$

Post-inflation, the expansion of the universe is first dominated by radiation with $H \sim 1/a^2$, and then by matter with $H \sim 1/a^{3/2}$. Even though the majority of the time is in the matter-dominated era, the vast majority of the expansion takes place in the radiation dominated era when energy densities were much higher. So we write $H_{\rm inf}/H_0 \sim (1 + z_{\rm inf})^2$. We then have

$$e^N > \left(\frac{H_{\rm inf}}{H_0} \right)^{1/2} = z_{\rm inf}$$

It remains to specify $H_{\rm inf}$ or, equivalently, $z_{\rm inf}$.

We don't currently know $H_{\rm inf}$. (We will briefly mention a way in which this can be measured in future experiments in Section 3.5.) However, as we will learn in Section 2, we understand the early universe very well back to redshifts of $z \sim 10^8 - 10^9$. Moreover, we're fairly confident that we know what's going on back to redshifts of $z \sim 10^{15}$ since

this is where we can trust the particle physics of the Standard Model. The general expectation is that inflation took place at a time before this, or

$$z_{\rm inf} > 10^{15} \quad \Rightarrow \quad N > 35$$

Recall that $H_0 \approx 10^{-18}$ s^{-1}, so if inflation took place at $z \approx 10^{15}$ then the Hubble scale during inflation was $H_{\rm inf} = 10^{12}\ s^{-1}$. In this case, inflation lasted a mere $T \sim 10^{-11}\ s$. These are roughly the time scales of processes that happen in modern particle colliders.

Many models posit that inflation took place much earlier than this, at an epoch where the early universe is getting close to Planckian energy scales. A common suggestion is

$$z_{\rm inf} \sim 10^{27} \quad \Rightarrow \quad N > 62$$

in which case $H_{\rm inf} \sim 10^{36}$ s^{-1} and $T \sim 10^{-35}$ s. This is an extraordinarily short time scale, and corresponds to energies way beyond anything we have observed in our puny experiments on Earth.

Most textbooks will quote around 60 e-foldings as necessary. For now, the take-away message is that, while there are compelling reasons to believe that inflation happened, there is still much we don't know about the process including the scale $H_{\rm inf}$ at which it occurred.

1.5.3 The Inflaton Field

Our theories of fundamental physics are written in terms of fields. These are objects which vary in space and time. The examples you've met so far are the electric and magnetic fields $\mathbf{E}(\mathbf{x}, t)$ and $\mathbf{B}(\mathbf{x}, t)$.

The simplest (and, so far, the only!) way to implement a transient, inflationary phase in the early universe is to posit the existence of a new field, usually referred to as the *inflaton*, $\phi(\mathbf{x}, t)$. This is a "scalar field", meaning that it doesn't have any internal degrees of freedom. (In contrast, the electric and magnetic fields are both vectors.)

The dynamics of this scalar field are best described using an action principle. In particle mechanics, the action is an integral over time. But for fields, the action is an integral over space and time. We'll first describe this action in flat space, and subsequently generalise it to the expanding FRW universe.

In Minkowski spacetime, the action takes the form

$$S = \int d^3x\, dt \ \left[\frac{1}{2}\dot{\phi}^2 - \frac{c^2}{2}\nabla\phi\cdot\nabla\phi - V(\phi)\right] \tag{1.78}$$

Here $V(\phi)$ is a potential. Different potentials describe different physical theories. We do not yet know the form of the inflationary potential, but it turns out that many do the basic job. (More detailed observations do put constraints on the form the potential can take as we will see in Section 3.5.) Later, when we come to solve the equations of motion, we will work with the simplest possible potential

$$V(\phi) = \frac{1}{2}m^2\phi^2 \tag{1.79}$$

The action (1.78) is then the field theory version of the harmonic oscillator. In the language of quantum field theory, m is called the *mass* of the field. (It is indeed the mass of a particles that arise when the field is quantised.)

The equations of motion for ϕ follow from the principle of least action. If we vary $\phi \to \phi + \delta\phi$, then the action changes as

$$\begin{aligned}\delta S &= \int d^3x\, dt \ \left[\dot{\phi}\,\delta\dot{\phi} - c^2\nabla\phi\cdot\nabla\delta\phi - \frac{\partial V}{\partial\phi}\delta\phi\right] \\ &= \int d^3x\, dt \ \left[-\ddot{\phi} + c^2\nabla^2\phi - \frac{\partial V}{\partial\phi}\right]\delta\phi\end{aligned}$$

where, in the second line, we have integrated by parts and discarded the boundary terms. Insisting that $\delta S = 0$ for all variations $\delta\phi$ gives the equation of motion

$$\ddot{\phi} - c^2\nabla^2\phi + \frac{\partial V}{\partial\phi} = 0$$

This is known as the *Klein-Gordon equation*. It has the important property that it is Lorentz covariant.

We want to generalise the action (1.78) to describe a scalar field in a homogenous and isotropic FRW universe. For simplicity, we restrict to the case of a $k = 0$ flat universe. This is a little bit unsatisfactory since we're invoking inflation in part to explain the flatness of space. However, it will allow us to keep the mathematics simple, without the need to understand the full structure of fields in curved spacetime. Hopefully, by the end you will have enough intuition for how scalar fields behave to understand that they will, indeed, do the promised job of driving the universe to become spatially flat.

In flat space, the FRW metric is simply

$$ds^2 = -c^2 dt^2 + a^2(t)\, d\mathbf{x}^2$$

The scale factor $a(t)$ changes the spatial distances. This results in two changes to the action (1.78): one in the integration over space, and the other in the spatial derivatives. We now have

$$S = \int d^3x\, dt\; a^3(t) \left[\frac{1}{2}\dot{\phi}^2 - \frac{c^2}{2a^2(t)} \nabla\phi \cdot \nabla\phi - V(\phi) \right] \tag{1.80}$$

Before we compute the equation of motion for ϕ, we first make a simplification: because we're only interested in spatially homogeneous solutions we may as well look at fields which are constant in space, so $\nabla\phi = 0$ and $\phi(\mathbf{x}, t) = \phi(t)$. We then have

$$S = \int d^3x\, dt\; a^3(t) \left[\frac{1}{2}\dot{\phi}^2 - V(\phi) \right] \tag{1.81}$$

Varying the action now gives

$$\delta S = \int d^3x\, dt\; a^3(t) \left[\dot{\phi}\, \delta\dot{\phi} - \frac{\partial V}{\partial \phi}\delta\phi \right] = \int d^3x\, dt\; \left[-\frac{d}{dt}\left(a^3 \dot{\phi} \right) - a^3 \frac{\partial V}{\partial \phi} \right] \delta\phi$$

Insisting that $\delta S = 0$ for all $\delta\phi$ again gives the equation of motion, but now there is an extra term because, after integration by parts, the time derivative also hits the scale factor $a(t)$. The equation of motion in an expanding universe is therefore

$$\ddot{\phi} + 3H\dot{\phi} + \frac{\partial V}{\partial \phi} = 0 \tag{1.82}$$

In the analogy with the harmonic oscillator, the extra term $3H\dot{\phi}$ looks like a friction term. It is sometimes referred to as *Hubble friction* or *Hubble drag*.

We also need to understand the energy density $\rho_{\rm inf} \equiv \rho_\rho$ associated to the inflaton field ϕ since this will determine the evolution of $a(t)$ through the Friedmann equation. There is a canonical way to compute this (through the stress-energy tensor) but the answer turns out to be what you would naively guess given the action (1.81), namely

$$\rho_\phi = \frac{1}{2}\dot{\phi}^2 + V(\phi) \tag{1.83}$$

The resulting Friedmann equation is then

$$H^2 = \frac{8\pi G}{3c^2} \left(\frac{1}{2}\dot{\phi}^2 + V(\phi) \right) \tag{1.84}$$

We will shortly solve the coupled equations (1.82) and (1.84). First we can ask: what kind of fluid is the inflaton field? To answer this, we need to determine the pressure. This follows straightforwardly by looking at

$$\dot{\rho}_\phi = \left(\ddot{\phi} + \frac{\partial V}{\partial \phi}\right)\dot{\phi} = -3H\dot{\phi}^2$$

Comparing to the continuity equation (1.39), $\dot{\rho} + 3H(\rho + P) = 0$, we see that the pressure must be

$$P_\phi = \frac{1}{2}\dot{\phi}^2 - V(\phi) \tag{1.85}$$

Clearly, this doesn't fit into our usual classification of fluids with $P = w\rho$ for some constant w. Instead, we have something more dynamical and interesting on our hands.

Slow Roll Solutions

We want to solve the coupled equations (1.82) and (1.84). In particular, we're looking for solutions which involve an inflationary phase. Taking the time derivative of (1.84), we have

$$2H\left(\frac{\ddot{a}}{a} - H^2\right) = \frac{8\pi G}{3c^2}\left(\ddot{\phi} + \frac{\partial V}{\partial \phi}\right)\dot{\phi} = -\frac{8\pi G}{c^2}H\dot{\phi}^2$$

where, in the second equality, we have used (1.82). Rearranging gives

$$\frac{\ddot{a}}{a} = -\frac{8\pi G}{3c^2}\left(\dot{\phi}^2 - V(\phi)\right)$$

which we recognise as the Raychaudhuri equation (1.52). We see that we get an inflationary phase only when the potential energy dominates the kinetic energy, $V(\phi) > \dot{\phi}^2$. Indeed, in the limit that $V(\phi) \gg \dot{\phi}^2$, the relationship between the energy (1.83) and pressure (1.85) becomes $P_\phi \approx -\rho_\phi$, which mimics dark energy.

Now we can get some idea for the set-up. We start with a scalar field sitting high on some potential, as shown in Figure 26 with $\dot{\phi}$ small. This will give rise to inflation. As the scalar rolls down the potential, it will pick up kinetic energy and we will exit the inflationary phase. The presence of the Hubble friction term in (1.82) means that the scalar can ultimately come to rest, rather than eternally oscillating backwards and forwards.

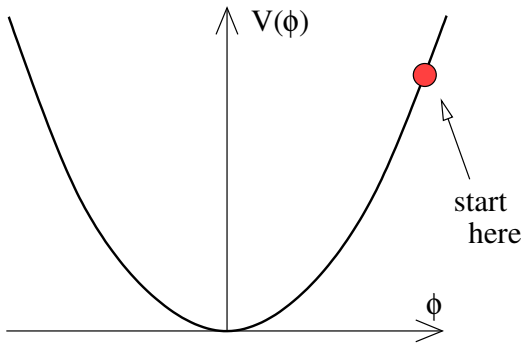

Figure 26: The inflationary scalar rolling down the potential $V(\phi)$.

Let's put some equations on these words. We assume that $V(\phi) \gg \frac{1}{2}\dot{\phi}^2$, a requirement that is sometimes called the *slow-roll condition*. The Friedmann equation (1.84) then becomes

$$H^2 \approx \frac{8\pi G}{3c^2} V(\phi) \tag{1.86}$$

Furthermore, if inflation is to last a suitably long time, it's important that the scalar does not rapidly gain speed. This can be achieved if the Hubble friction term dominates in equation (1.82), so that $\ddot{\phi} \ll H\dot{\phi}$. In the context of the harmonic oscillator, this is the over-damped regime. The equation of motion is then

$$3H\dot{\phi} \approx -\frac{\partial V}{\partial \phi} \tag{1.87}$$

These are now straightforward to solve. For concreteness, we work with the quadratic potential $V = \frac{1}{2}m^2\phi^2$. Then the solutions to (1.86) and (1.87) are

$$H = \alpha\phi \quad \text{and} \quad \dot{\phi} = -\frac{m^2}{3\alpha} \quad \text{with } \alpha^2 = \frac{4\pi G m^2}{3c^2}$$

Integrating the second equation gives

$$\phi(t) = \phi_0 - \frac{m^2}{3\alpha}t$$

where we have taken the scalar field to start at some initial value ϕ_0 at $t = 0$. We can now easily integrate the $H = \alpha\phi$ equation to get an expression for the scale factor,

$$a(t) = a(0)\, \exp\left[\frac{2\pi G}{c^2}(\phi_0^2 - \phi(t)^2)\right] \tag{1.88}$$

This is a quasi-de Sitter phase of almost exponential expansion.

This solution remains valid provided that the condition $V(\phi) \gg \dot{\phi}^2$ is obeyed. The space will cease to inflate when $V(\phi) \approx \dot{\phi}^2$, which occurs when $\phi^2(t_{\rm end}) \approx 2m^2/(3\alpha)^2$. By this time, the universe will have expanded by a factor of

$$\frac{a(t_{\rm end})}{a(0)} \approx \exp\left[\frac{2\pi G\phi_0^2}{c^2} - \frac{1}{3}\right]$$

We see that, by starting the scalar field higher up the potential, we can generate an exponentially large expansion.

1.5.4 Further Topics

There is much more to say about the physics of inflation. Here we briefly discuss a few important topics, some of which are fairly well understood, and some of which remain mysterious or problematic.

Reheating

By the end of inflation, the universe is left flat but devoid of any matter or radiation. For this to be a realistic mechanism, we must find a way to transfer energy from the inflaton field into more traditional forms of matter. This turns out to be fairly straightforward, although we are a long way from a detailed understanding of the process. Roughly speaking, if the inflaton field is coupled to other fields in nature, then these will be excited as the inflaton oscillates around the minimum of its potential. This process is known as *reheating*. Afterwards, the standard hot Big Bang cosmology can start.

Dark Energy or Cosmological Constant?

Inflation is a period of dynamically driven, temporary, cosmic acceleration in the very early universe. Yet, as we have seen, the universe is presently entering a second stage of comic acceleration. How do we know that this too isn't driven by some underlying dynamics and will, again, turn out to be temporary? The answer is: we don't. It is not difficult to cook up a mathematical model in which the cosmological constant is set to zero by hand and the current acceleration is driven using some scalar field. Such models go by the unhelpful name of *quintessence*.

Quintessence models are poorly motivated and do nothing to solve the fine-tuning problems of the cosmological constant. In fact, they are worse. First, we have to set the genuine cosmological constant to zero (and we have no reason to do so) and then we have to introduce a new scalar field which, to give the observed acceleration, must have an astonishingly small mass of order $m \sim 10^{-33}\ eV$.

Such models look arbitrary and absurd. And yet, given our manifest ignorance about the cosmological constant, it is perhaps best to keep a mildly open mind. The smoking gun would be to measure an equation of state $P = w\rho$ for the present day dark energy which differs from $w = -1$.

Initial Conditions

For the idea of inflation to fly, we must start with the scalar field sitting at some point high up the potential. It is natural to ask: how did it get there?

One possibility is that the initial value of the scalar field varies in space. The regions where the scalar are biggest then inflate the most, and all traces of the other regions are washed away beyond the horizon. These kind of ideas raise some thorny issues about the nature of probabilities in an inflationary universe (or multiverse) and are poorly understood. Needless to say, it seems very difficult to test such ideas experimentally.

A More Microscopic Underpinning?

Usually when we introduce a scalar field in physics, it is an approximation to something deeper going on underneath. For example, there is a simple theory of superconductivity, due to Landau and Ginsburg, which invokes a scalar field coupled to the electromagnetic field. This theory makes little attempt to justify the existence of the scalar field. Only later was a more microscopic theory of superconductivity developed — so-called BCS theory — in which the scalar field emerges from bound pairs of electrons. Many further examples, in which scalar fields are invoked to describe everything from water to magnets, can be found in the lectures on Statistical Field Theory.

This raises a question: is the scalar field description of inflation an approximation to something deeper going on underneath? We don't know the answer to this.

Quantum Fluctuations

Although inflation was first introduced to solve the flatness and horizon problems, its greatest triumph lies elsewhere. As the scalar field rolls down the potential, it suffers small quantum fluctuations. These fluctuations are swept up in the expansion of the universe and stretched across the sky where, it is thought, they provide the seeds for the subsequent formation of structure in the universe. These fluctuations are responsible for the hot and cold spots in the CMB which, in turn, determine where matter clumps and galaxies form. In Section 3.5 we will look more closely at this bold idea.