PHYSICS 156 - A COSMOLOGY WORKBOOK

Lloyd Knox
University of California, Davis

LibreTexts™

University of California, Davis

Physics 156 - A Cosmology Workbook

Lloyd Knox

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden born from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.

LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by NICE CXone Expert and was supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org or visit our main website at https://LibreTexts.org.

This text was compiled on 06/12/2026

TABLE OF CONTENTS

Licensing

About the Authors

1: Workbook

- 1.1: Overview
- 1.2: Spacetime Geometry
- 1.3: Redshifts
- 1.4: Spatially Homogeneous and Isotropic Spacetimes
- 1.5: Euclidean Geometry
- 1.6: Distances as Determined by Standard Candles
- 1.7: The Distance-Redshift Relation
- 1.8: Dynamics of the Expansion
- 1.9: A Newtonian Homogeneous Expanding Universe
- 1.10: The Friedmann Equation
- 1.11: Particle Kinematics in an Expanding Universe - Newtonian Analysis
- 1.12: The Evolution of Mass-Energy Density and a First Glance at the Contents of the Cosmos
- 1.13: Energy and Momentum Conservation
- 1.14: Pressure and Energy Density Evolution
- 1.15: Distance and Magnitude
- 1.16: Parallax, Cepheid Variables, Supernovae, and Distance Measurement
- 1.17: Cosmological Data Analysis
- 1.18: The Early Universe
- 1.19: Equilibrium Statistical Mechanics
  - 1.19.1: Chapter 19 footnotes
- 1.20: Equilibrium Particle Abundances
- 1.21: Hot and Cold Relics of the Big Bang
- 1.22: Overview of Thermal History
- 1.23: Big Bang Nucleosynthesis - Predictions
- 1.24: Big Bang Nucleosynthesis - Observations
- 1.25: Introduction to the Cosmic Microwave Background
- 1.26: The Spectrum of the CMB
- 1.27: Cosmic Microwave Background Anisotropies
- 1.28: Solving the Wave Equation with Fourier Transforms
- 1.29: The First Few Hundred Thousand Years- The Dynamics of the Primordial Plasma
- 1.30: Structure Formation
- 1.31: Galaxy Formation
- 1.32: Euclidean Geometry
- 1.33: Curvature
- 1.34: Galilean Relativity
- 1.35: Einstein Relativity
- 1.36: The Simplest Expanding Spacetime
- 1.37: Redshifts

Index

Glossary

Detailed Licensing

Detailed Licensing

Licensing

*A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.*

About the Authors

Ryan Cooke

Ryan Cooke is an Associate Professor of Physics at Durham University, and a Royal Society University Research Fellow. His primary research interests include primordial nucleosynthesis, the first stars, cosmology and fundamental physics. For more information about Professor Cooke's research, and contact details, please visit this website. He is the author of the chapter on Big Bang Nucleosynthesis measurements in A Cosmology Workbook.

Lloyd Knox

Lloyd Knox is a Professor of Physics and Astronomy at UC Davis, where he has been on the faculty since 2000. In 2012 he was elected as a Fellow of the American Physical Society with a citation that reads in part, "For motivating major observations (WMAP and Planck), developing widely-used data analysis tools, providing insightful interpretations of data, and calculating the impact of astrophysical processes on the microwave sky." You can read more about Lloyd Knox here. He is the chief editor of A Cosmology Workbook and the author of most of its chapters.

Ethan Nadler

Ethan Nadler is a joint postdoctoral fellow at the Carnegie Observatories and USC. His research combines cosmological simulations, particle theory, and observations of the smallest galaxies and cosmic structures to study the microphysical nature of dark matter. He also works at the interface of data and dark matter theory with collaborations including the Dark Energy Survey, Dark Energy Science Collaboration, and Satellites Around Galactic Analogs Survey. He is the author of the chapters on Structure Formation and Galaxy Formation in A Cosmology Workbook.

CHAPTER OVERVIEW

1: Workbook

It has been quite a century for our understanding of the cosmos. As I write these words at the beginning of 2017 it was just over 100 years ago, in November 1915, that Albert Einstein finished development of his general theory of relativity. Among many other things, this theory provided the proper context for interpreting Edwin Hubble's distance-redshift law, published in 1929, as due to the expansion of the Universe. In the 40s, Gamow and Alpher speculated that the dense conditions that must have existed earlier in this expanding universe could provide another site, in addition to the cores of stars, for the fusion of light elements to heavier ones. In fact, to avoid over-production of the elements, this earlier, denser phase would have to be very hot. In the 1960s Bell Labs scientists accidentally stumbled upon the thermal radiation left over from that heat, that exists in the current epoch as a nearly uniform microwave glow. With that discovery, the idea that the universe used to be hot, dense, and expanding very rapidly became the dominant cosmological paradigm known as the "Big Bang."

The subsequent fifty years of the past century saw much progress as well. We now know that we do *not* know what constitutes 95% of the mass/energy in the universe. Only 5% of the mass/energy is composed of constituents in the particle physicist's standard model. Most of the rest is "dark energy" which smoothly fills the universe and dilutes only slowly, if at all, as the universe expands. The rest is "dark matter" that, like George Lucas's mystical Force, "pervades us and binds the galaxy together." Measurements of light element abundances, combined with modern, precision, version's of Gamow and Alpher's big bang nucleosynthesis calculations, give us confidence we understand the expansion back to an epoch when the presently observable universe was 10^{27} times smaller in volume than it is now. A speculative theory, known as cosmic inflation, has met with much empirical success, giving us some level of confidence we may understand something about events at yet higher densities and even earlier times.

In this quarter-long course we will at least touch upon all the topics in the above two paragraphs. We will learn how to think about the expanding universe using concepts from Einstein's theory of general relativity. We will use Newtonian gravity to derive the dynamical equations that relate the expansion rate to the matter content of the universe. Connecting the expansion dynamics to observables such as luminosity distances and redshifts, we will see how astronomers use observations to probe these dynamics, and thereby the contents of the cosmos, including the mysterious dark energy.

We will introduce some basic results of kinetic theory to understand why big bang nucleosynthesis leads to atomic matter that is, by mass, about 25% Hydrogen, 75% Helium with only trace amounts of heavier elements. We'll use this kinetic theory, applied to atomic rather than nuclear reactions, to explore perhaps the most informative cosmological observable: the cosmic microwave background. Finally, we will study how an early epoch of inflationary expansion, driven by an exotic material with negative pressure, can explain some of the otherwise puzzling features of the observed universe.

1.1: Overview
1.2: Spacetime Geometry
1.3: Redshifts
1.4: Spatially Homogeneous and Isotropic Spacetimes
1.5: Euclidean Geometry
1.6: Distances as Determined by Standard Candles
1.7: The Distance-Redshift Relation
1.8: Dynamics of the Expansion
1.9: A Newtonian Homogeneous Expanding Universe
1.10: The Friedmann Equation
1.11: Particle Kinematics in an Expanding Universe - Newtonian Analysis
1.12: The Evolution of Mass-Energy Density and a First Glance at the Contents of the Cosmos
1.13: Energy and Momentum Conservation
1.14: Pressure and Energy Density Evolution
1.15: Distance and Magnitude
1.16: Parallax, Cepheid Variables, Supernovae, and Distance Measurement
1.17: Cosmological Data Analysis

1.18: The Early Universe
1.19: Equilibrium Statistical Mechanics
1.19.1: Chapter 19 footnotes
1.20: Equilibrium Particle Abundances
1.21: Hot and Cold Relics of the Big Bang
1.22: Overview of Thermal History
1.23: Big Bang Nucleosynthesis - Predictions
1.24: Big Bang Nucleosynthesis - Observations
1.25: Introduction to the Cosmic Microwave Background
1.26: The Spectrum of the CMB
1.27: Cosmic Microwave Background Anisotropies
1.28: Solving the Wave Equation with Fourier Transforms
1.29: The First Few Hundred Thousand Years- The Dynamics of the Primordial Plasma
1.30: Structure Formation
1.31: Galaxy Formation
1.32: Euclidean Geometry
1.33: Curvature
1.34: Galilean Relativity
1.35: Einstein Relativity
1.36: The Simplest Expanding Spacetime
1.37: Redshifts

Thumbnail: This modification of the Flammarion Engraving is an enigmatic woodcut by an unknown artist. The woodcut depicts a man peering through the Earth's atmosphere as if it were a curtain to look at the inner workings of the universe. The original caption below the picture (not included here) translated to: "A medieval missionary tells that he has found the point where heaven and Earth meet...".

Contributor

- Lloyd Knox (UC Davis)

This page titled 1: Workbook is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.1: Overview

Cosmologists speak with a high degree of confidence about conditions that existed billions of years ago when the universe was quite different from how we find it today: 10^9 times hotter than today, over 10^{30} times denser, and much much smoother, with variations in density from one place to another only as large as one part in 100,000. We claim to know the composition of the universe at this early time, dominated almost entirely by thermal distributions of photons and subatomic particles called neutrinos. We know in detail many aspects of the evolutionary process that connects this early universe to the current one. Our models of this evolution have been highly predictive and enormously successful.

In this chapter we provide an overview of our subject, broken into two parts. The first is focused on the discovery of the expansion of the universe in 1929, and the theoretical context for this discovery, which is given by Einstein's general theory of relativity (GR). The second is on the implications of this expansion for the early history of the universe, and relics from that period observable today: the cosmic microwave background and the lightest chemical elements. The consistency of such observations with theoretical predictions is why we speak confidently about such early times, approximately 14 billion years in our past.

Overview Part I: The Expansion of Space and the Contents of the Universe

Space is not what you think it is. It can curve and it can expand over time. We can observe the consequences of this expansion over time, and from these observations infer some knowledge of the Universe's contents.

Newton-Maxwell Incompatibility and Einstein's Theory

It would be difficult to overstate the impact that Einstein’s 1915 General Theory of Relativity has had on the field of cosmology. So we begin our review with some discussion of the General Theory, and its origin in conflicts between Newtonian physics and Maxwell’s theory of electric and magnetic fields. We end our discussion of Einstein’s theory with its prediction that the universe must be either expanding or contracting.

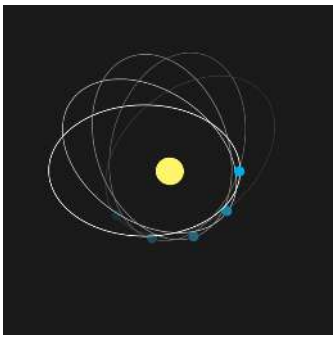
Newton’s laws of motion and gravitation have amazing explanatory powers. Relatively simple laws describe, almost perfectly, the motions of the planets and the moon, as well as the motions of bodies here on Earth -- at least at speeds much lower than the speed of light. The discovery of the planet Neptune provides us with an example of their predictive power. The discovery began with calculations by Urbain Le Verrier. Using Newton’s theory, he was able to explain the observed motion of Uranus only if he posited the existence of a planet with a particular orbit, an extra planet beyond those known at the time. Without the gravitational pull of this not-yet-seen planet, Newton's theory could not account for the motion of Uranus. Following up on his prediction, made public in 1846, Johannes Gottfried Galle looked for a new planet where Le Verrier said it was to be found and, indeed, there it was, what we now call Neptune. The time from prediction to confirmation was less than a month.

Less than 20 years after the discovery of Neptune, a triumph of the Newtonian theory, came a great inductive synthesis: Maxwell’s theory of electric and magnetic fields. The experiments that led to this synthesis, and the synthesis itself, have enabled the development of a great range of technologies we now take for granted such as electric motors, radio, television, cellular communication networks and microwave ovens. More important for our subject, they also led to radical changes to our conception of space and time.

These radical changes arose from conflicts between the synthesis of Maxwell with that of Newton. For example, in the Newtonian theory velocities add: if A sees B move at speed v to the west, and B sees C moving at speed v to the west relative to B, then A sees C moving at speed v+v = 2v to the west. But one solution to the Maxwell equations is a propagating disturbance in the electromagnetic fields that travels with a fixed speed of about 300,000 km/sec. Without modification, the Maxwell equations predict that both A and B would see electromagnetic wave C moving away from them at a speed of 300,000 km/sec, violating the velocity addition rule that one can derive from Newtonian concepts of space and time.

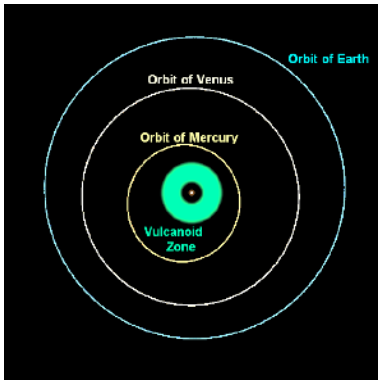
Einstein’s solution to these inconsistencies includes an abandonment of Newtonian concepts of space and time. This abandonment, and discovery of the replacement principles consistent with the Maxwell theory, happened over a considerable amount of time. A solution valid in the absence of gravitation came out first in 1905, with Einstein’s paper titled (in translation from German) “On the Electrodynamics of Moving Bodies.” His effort to reconcile gravitational theory with Maxwell’s theory did not fully come together until November of 1915, with a series of lectures in Berlin where he presented his General Theory of Relativity.

One indicator that Einstein was on the right track was his realization, in September 1915, that his theory provided an explanation for a longstanding problem in solar system dynamics known as the anomalous perihelion precession of Mercury. Given Newtonian theory, and an absence of other planets, Mercury would orbit the Sun in an ellipse shape. However, the influence of the other



planets is to make Mercury follow almost an ellipse, in a pattern that is well approximated as that of a slowly rotating ellipse. One way of expressing this rotation is to say how rapidly the location of closest approach, called perihelion, is rotating around the Sun. Mercury's perihelion precession is quite slow. In fact, it's less than one degree per century. More precisely, it's 575 seconds of arc per century, where a second of arc is a degree divided by 3,600 (just as a second is an hour divided by 3,600).

Urbain Le Verrier, following his success with Neptune, took up the question of this motion of Mercury: could the perihelion precession be understood as resulting from the pulls on Mercury from the other planets. He found that he could ascribe about 532 seconds of arc to the other planets, but not the entire 575. There is an additional, unexplained ("anomalous") precession of 43 seconds of arc per century. Le Verrier, of course, knew how to handle situations like this. He proposed that this motion is caused by a not-yet-discovered planet. This planet was proposed to have an orbit closer to the Sun than Mercury's and eventually had the name Vulcan.



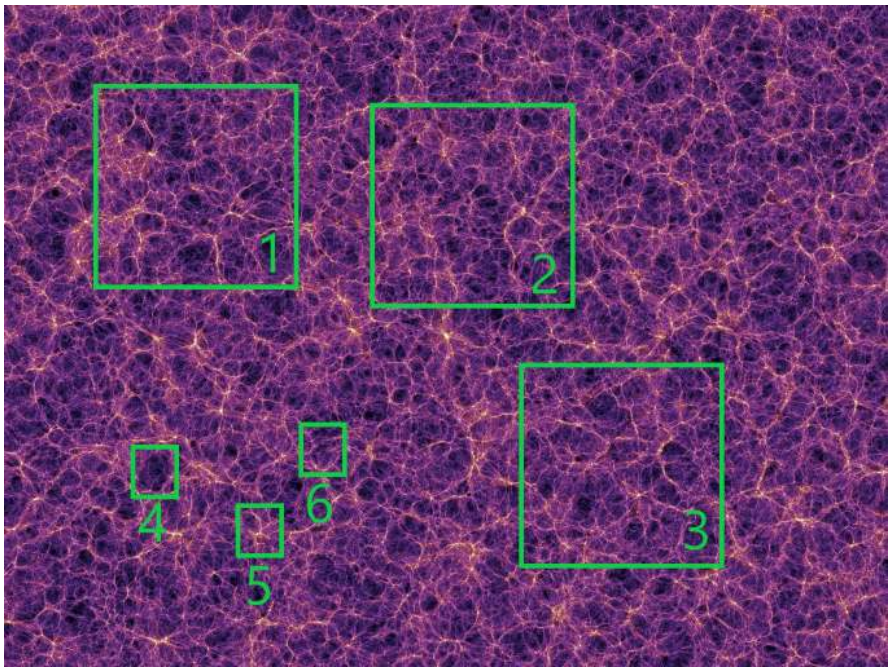
But "Vulcan, Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune, Pluto" is probably not the list of planets you learned in elementary school! Unlike the success with Neptune, Newtonian gravity was not going to be vindicated by discovery of another predicted planet. Rather than an unaccounted-for planet, the anomalous precession of Mercury, we now know, as Einstein figured out in September of 1915, is due to a failure of Newton's theory of gravity. At slow speeds and for weak gravitational fields the Newtonian theory is an excellent approximation to Einstein's theory -- so the largest errors in the Newtonian theory show up for the fastest-moving planet orbiting closest to the Sun.

Amazingly, thinking about a theory that explains experiments with electricity and magnets, and trying to reconcile it with Newton's laws of gravitation and motion, had led to a solution to this decades-old problem in solar system dynamics. The theory has gone on from success to success since that time, most recently with the detection of gravitational waves first reported in 2016. It is of practical importance in the daily lives of many of us: GPS software written based on Newtonian theory rather than Einstein's theory would be completely useless.

The Expansion of Space

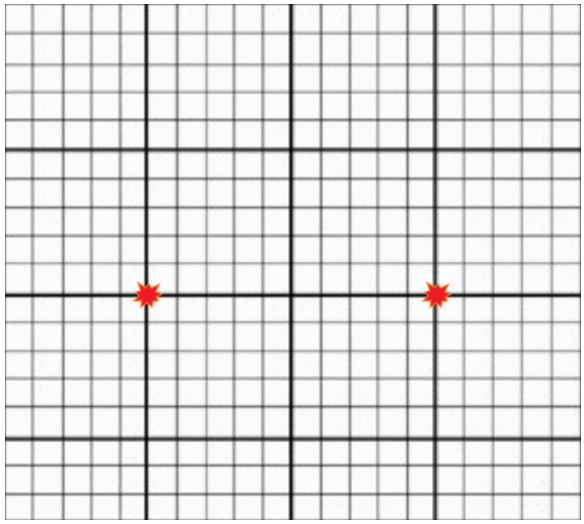
More important to our subject, Einstein's theory allowed for better-informed speculation about the history of the universe as a whole. In the years following Einstein's November 1915 series of lectures, a number of theoreticians calculated solutions of the Einstein equations for highly-idealized models of the universe. The Einstein field equations are extremely difficult to solve in generality. The first attempts at solving these equations for the universe as a whole thus involved extreme idealization. They used what you might call "the most spherical cow approximation of all time." They approximated the whole universe as completely homogeneous; i.e., absolutely the same everywhere.

We now know that on very large scales, this is a good approximation to our actual universe. To illustrate what we mean by homogeneity being a good approximation on very large scales, we have the figure below which shows a slice from a large-volume simulation of the large-scale structure of our universe. In these images, brighter regions are denser regions. The image has two sets of sub-boxes: large ones and small ones. We can see that the universe appears different in the small boxes. Box 4 is under dense, Box 5 is over dense, and Box 6 is about average. If we look at the larger boxes, the universe appears more homogeneous. Each box looks about the same. This is the sense by which we mean that on large scales the universe is highly homogeneous.



On large scales the Universe is highly homogeneous. The large boxes (boxes 1, 2, and 3) are about 200 Mpc across (that's about 600 million light years). No matter where you put down such a large box, the contents look similar. For the smaller boxes this is not the case. Images are from the Millenium Simulation.

Gif of an expanding grid (below) **Image by Alex Eisner**



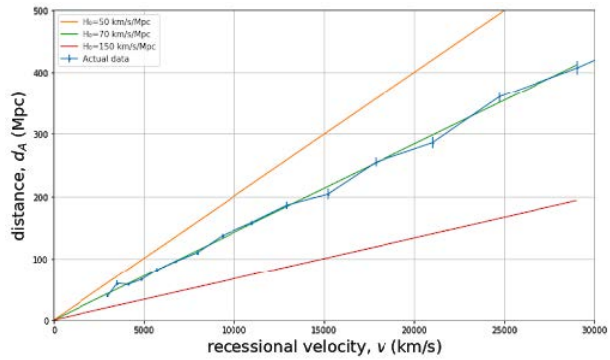
D vs. z: Low z

In 1929, Edwin Hubble made an important observation by measuring distances to various galaxies and by measuring their "redshifts." Hubble inferred distances to galaxies by using standard candles, which are objects with predictable luminosities. Since farther away objects appear dimmer, one can predict distances by comparing the object's expected luminosity with how bright it appears. The redshift of a galaxy, which cosmologists label as "z", tells us how the wavelength of light has shifted during its propagation. Mathematically, $1+z$ is equal to the ratio of the wavelength of observed light to the wavelength of emitted light: $(1+z = \lambda_{\text{observed}}/\lambda_{\text{emitted}})$. At least for $z << 1$, one can think of this as telling us how fast the galaxy is moving away from us according to the Doppler effect: a higher redshift indicates a larger velocity with relationship $v = cz$ with c the speed of light. If a galaxy were instead moving towards us, its light would appear blueshifted. Hubble found that not only were nearly all the galaxies redshifted, but there was a linear relationship between the galaxies' distances and redshifts. This is represented by Hubble's law, v=H_0d.

Hubble's observations, which were evidence for this simple law, had profound consequences for our understanding of the cosmos: they indicated that the universe was expanding.

To understand this, take a look at the image of an expanding grid. Each of the two red points is “stationary”; i.e., they each have a specific defined location on the grid and do not move from that location. However, as the grid itself expands, the distance between the two points grows, and they appear to move away from each other. If you lived anywhere on this expanding grid, you would see all other points moving away from you. It's easiest to see this is true for the central location on the grid. Try placing your mouse over different points on the grid and following them as they expand. You will notice that points farther from the center will move away faster than points close to the center. This is how Hubble’s law and the expansion of space works. Instead of viewing redshifts as being caused by a Doppler effect from galaxies moving through space, we will come to understand the redshift as the result of the ongoing creation of new space.

Hubble’s constant, H_0, tells us how fast the universe is expanding in the current epoch. It was first estimated to be about 500 km/s/Mpc, but after eliminating some gross systematic errors and obtaining more accurate measurements, we know this initial estimate was off, and not by any small amount. It's wrong by about a factor of 7! The exact value of the Hubble constant is actually somewhat controversial today, but everyone agrees it's somewhere between 66 and 75 km/s/Mpc. This means that on large length scales, where the approximation of a homogeneous universe becomes more accurate, about 70 kilometers of new space are created each second in every Megaparsec.



Hubble’s law shown for low redshifts, where we can think of the redshift as arising due to a Doppler effect and so v=cz. Blue line is actual data from supernovae, green line is a best-fit line for a Hubble constant of 70 km/s/Mpc. Orange and red lines show how this relationship would differ for other Hubble constants. Image by Adrianna Schroeder.

D vs. z: High z

As we will see, from Einstein’s theory of space and time we can expect the rate of expansion of space to change over time. The history of these changes to the expansion rate leave their impact on the distance-redshift relation if we trace it out to sufficiently large distances and redshifts. As we measure out to larger distances, the relationship is no longer governed by cz = H_0 d. Instead, we will show that the redshift tells us how much the universe has expanded since light left the object we are observing.

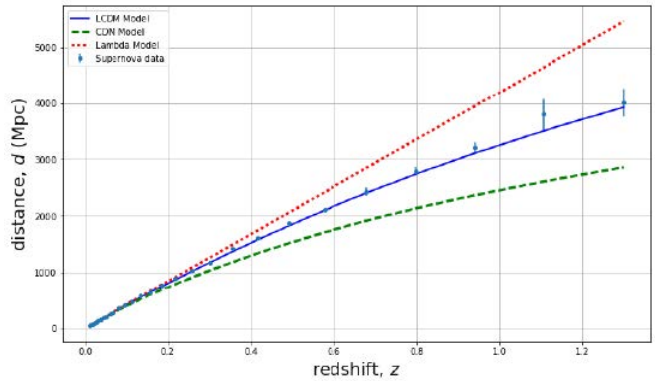
$$1 + z = \frac{\lambda_{observed}}{\lambda_{emitted}} = \frac{a_0}{a_e} \tag{1.1.1}$$

where a is the “scale factor” that parameterizes the expansion of space, a_0 is the scale factor today and a_e is the scale factor when the light was emitted. We can observe quasars so far away that the universe has expanded by a factor of 7 since the light left them that we are receiving now. For such a quasar we have a_0/a_e = 7 so the wavelength of light has been stretched by a factor of 7, and by definition of redshift z we have z = 6.

The distance from us to such a quasar depends on how long it took for the universe to expand by a factor of 7. If the expansion rate were slower over this time, then it would have taken longer, so the quasar must be further away. Measurements of distance vs. redshift are thus sensitive to the history of the expansion rate.

As we will see, how the expansion rate changes over time depends on what the universe is made out of. Therefore, studying D vs. z out to high distances and redshifts can help us determine the composition of the universe. We can see such measurements in Fig. B, together with some model curves. The models all have the same expansion rate today, H_0, but differ in the mix of different kinds of matter/energy in the universe. A model that’s purely non-relativistic matter, the green dashed line "CDM Model", does a very poor job of fitting the data. The data seem to require a contribution to the energy density that we call “vacuum energy” or “the cosmological constant.” The “Lambda CDM” or "LCDM" model has a mixture of this cosmological constant and non-relativistic

matter. The cosmological constant causes the expansion rate to accelerate. Thus, compared to the model without a cosmological constant (the "CDM Model"), the expansion rate in the past was slower; it thus took a longer time for the universe to expand by a factor of 1+z, and thus objects with a given z are at further distances. This acceleration of the expansion rate, discovered via D vs. z measurements at the end of the 20th century, was a great surprise to most cosmologists. Why there is a cosmological constant, or whether there is something else causing the acceleration, is one of the great mysteries of modern cosmology.



Redshift versus distance up to higher redshifts. The shape of the graph can tell us about the contents of the universe--current data are fit well by the “LCDM” model, which we will learn about later. Note: we use z here for our x axis rather than the recessional velocity v = cz. Although for low z, we can think of redshift as arising from a Doppler effect, that interpretation only makes sense for z << 1. More generally, as we will see, 1+z is the amount by which the universe expands since light left the object we are observing. Image by Adrianna Schroeder.

Overview Part II: The Hot Big Bang and its Relics

The expansion of the universe implies that it must have been much smaller, and much denser, in the past. If this is true, we should be able to see some consequences from the very high density period of the early universe. One such consequence is the relative abundances of light elements we see today. Most of the Helium, the second-most-common element in the Universe, was created when the expansion was less than a few minutes old. Trace amounts of other light elements were also created in this early period in a process we call Big Bang Nucleosynthesis (BBN). Observation of the abundances of these light elements can tell us about conditions at such early times. To achieve consistency between predictions and observations requires that the big bang was very hot, and thereby led to the prediction of what we call the cosmic microwave background. The discovery of this background in 1964 led to the establishment of the hot big bang model as the standard cosmological paradigm.

Our First Relic: Light Elements

In 1948, Gamow explored the very early universe as a source of elements heavier than hydrogen. He extrapolated Einstein’s theory of an expanding universe with certain assumptions about the composition of the universe and concluded that it was infinitely dense at a finite time in the past. He theorized that this early universe could be a prodigious source of heavier elements. If the universe began at infinitely high density and temperatures and underwent rapid expansion and cooling, atomic nuclei would form not too early, when high-energy radiation did not permit any nuclei to survive, and not too late, when temperatures were too low for the nuclear collisions to overcome the Coulomb repulsion, but at a just-right intermediate stage.

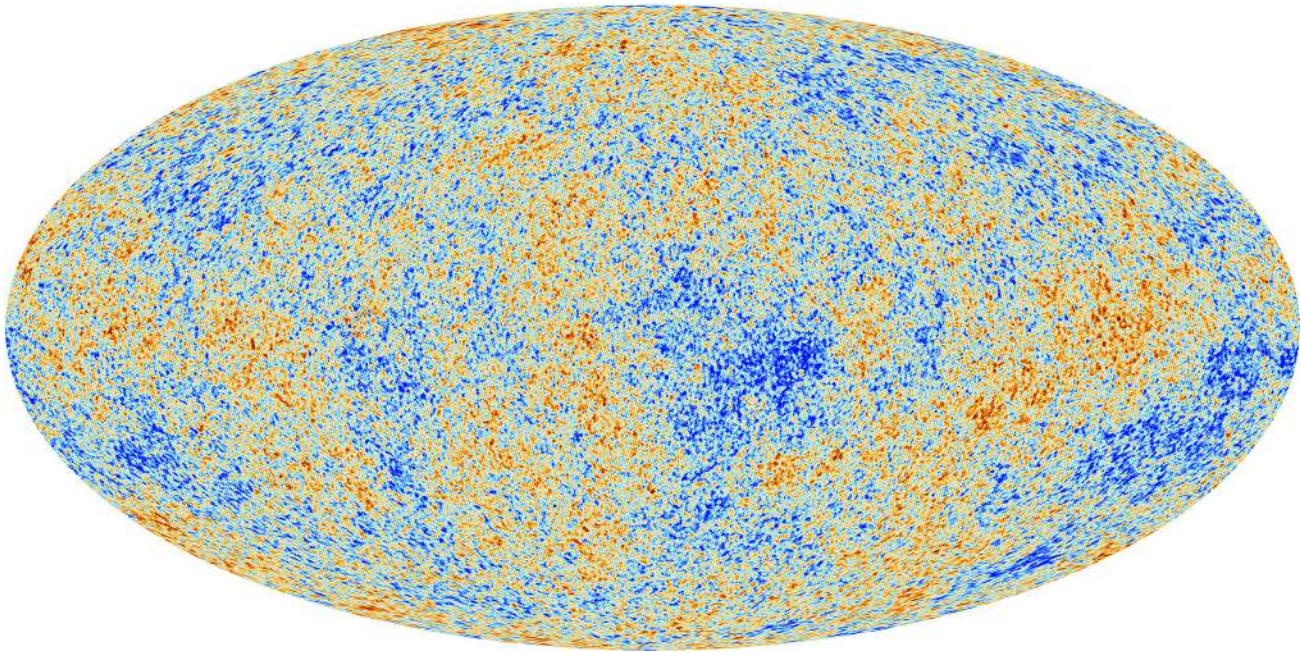
As we will see, almost all of the hydrogen and helium in the universe originated in the big bang. Further, those are the only elements to be produced in the big bang in anything beyond trace amounts. In our chapters on big bang nucleosnythesis we will explore the theory of the production of helium, and compare the predicted amounts of helium and trace amounts of deuterium and lithium with attempts to infer those abundances from observations.

Our Next Relic: The Cosmic Microwave Background

An important aspect of Gamow's work was that in order to avoid overproduction of helium and other heavy elements, the ratio of nucleons to photons at this just-right epoch had to be very small. Since the number of photons in black body radiation is proportional to temperature cubed, this means that the “Big Bang” had to be very, very hot. Pushing this chain of logic forward in a follow-up paper, Alpher and Herman theorized that we should see a background of heat and light from this period of high temperature and photon density. This background was discovered later in 1964 and is known as the Cosmic Microwave Background (CMB).

The CMB is an enormous gift from nature to cosmologists. By measuring both the spectrum of this light, and mapping its variation in intensity and polarization across the sky, we have a highly sensitive probe of conditions in the universe from times of thousands of years to hundreds of thousands of years. What's more, the conditions in the universe at this time were such that one can calculate, with great accuracy, the outcomes for the CMB spectrum, intensity, and polarization measurements, to be expected for some particular model of the cosmos. Physical systems with such extreme calculability and utility are rare. The CMB is to cosmology what the solar system was for science and Newtonian gravitational theory three hundred plus years ago.

The following image is a projection of intensity variations across the whole (curved) sky on to a flat map. To better understand how this relates to what we observe from Earth, explore the virtual CMB planetarium below it by clicking and dragging. This shows how the CMB would look over a tree-lined horizon, if human eyes were extremely sensitive to light at millimeter wavelengths.



As we will see, the consistency between the statistical properties of CMB anisotropy and polarization maps, and the predictions of the standard cosmological model, LCDM, are quite extraordinary. They give us some confidence that we know what we are talking about when we describe conditions in the universe when it is just a few hundred thousand years old and even younger.

Other Relics of the Big Bang: the Cosmic Neutrino Background and Dark Matter

The hot and dense conditions of the big bang not only led to thermal production of a background of photons (the CMB), but also a background of neutrinos, and possibly dark matter as well. According to our standard cosmological model, most of the mass/energy density of the universe was at one time (just prior to the epoch of Big Bang Nucleosythesis) in the Cosmic Neutrino Background (CNB). This background has yet to be directly detected, but has been detected via its gravitational influence on the photons.

To explain a long list of different observed phenomena, our standard cosmological model also includes a significant amount of non-relativistic matter that does not interact with atoms, nuclei, and light, that we call "dark matter." The amount of dark matter in a large representative sample of the universe appears to be about six times the density of matter made of out of atomic nuclei and electrons. To date we only know of this dark matter via its gravitational influence; it has not yet been directly detected. Dark matter might also be a thermal relic of the Big Bang.

Conclusion

The agreement between cosmological predictions and observations give us a high degree of confidence that our models are capturing some important and true things about the nature of the cosmos. However, we still don’t know what most of the universe is made of, or how it came into existence. There is work left to do!

Our measurements are constantly improving, and one never knows when a combination of precise measurements and predictions will reveal something new and interesting about the universe, or when a theoretical insight will resolve some long-standing puzzles

and shine a light for us to follow towards a more perfect understanding.

The adventure continues. Maybe you will join us in our quest!

This page titled 1.1: Overview is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.2: Spacetime Geometry

In this chapter we introduce the geometrical concept of a spacetime, how we describe it mathematically, and the relationship of the mathematical quantities to things we can measure with clocks and rulers. We begin with a spacetime we assume that you have studied before: the Minkowski spacetime of special relativity. After some review of special relativity, to get everyone oriented to the notation, we then generalize the spacetime slightly to one that is uniformly *expanding*. With additional assumptions we then calculate the age of this spacetime as a function of expansion rate, as well as the "past horizon."

The Invariant Distance

We can label spacetimes with coordinates; for example, we could label every point in a spacetime with one spatial dimension (a so-called 1+1-dimensional spacetime) with a t value and an x value. These coordinates are just labels, with no physical meaning on their own. Physical meaning comes via a rule that connects infinitesimally-separated pairs of points to measurements with clocks and rulers. More specifically, the rule gives the square of the "invariant distance" between infinitesimally-separated pairs of points, which we denote as ds^2. Before giving an example of such a rule, let's make completely precise the relationship between ds^2 and what can be measured with clocks and rulers.

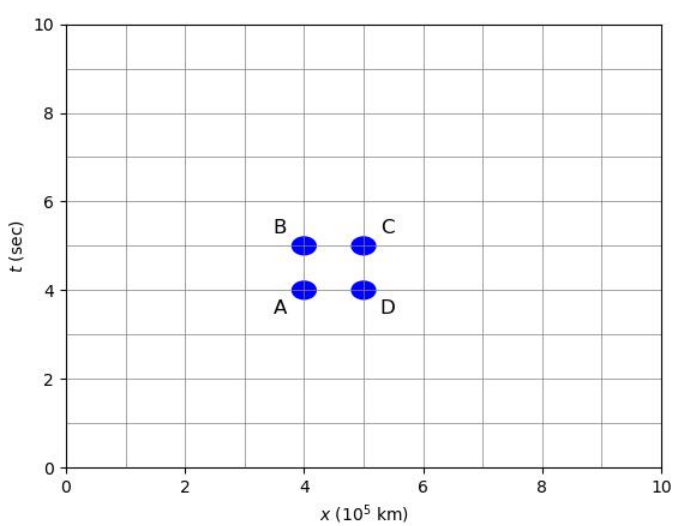
Note

The physical interpretation of ds^2 is as follows:

1. For $ds^2 < 0$ (which we call time-like separations), the time elapsed on a clock that travels between the two space-time points is $\sqrt{-ds^2/c^2}$; and
2. For $ds^2 > 0$ (which we call space-like separations), the length of a ruler at rest in the frame in which the two events are simultaneous, with an end on each of the two space-time points, is $\sqrt{ds^2}$.

For example, in a 1+1-dimensional Minkowski spacetime it is possible to label it such that the square of the invariant distance between a point labeled t, x and another point labeled $t + dt, x + dx$ is given by:

$$ds^2 = -c^2 dt^2 + dx^2. \tag{1.2.1}$$



Let's consider this rule for the points labeled A, B, C and D in the figure. In this coordinate system, the point labeled by A is also labeled by $t = 4$ seconds and $x = 4 \times 10^5$ km. Let's calculate the length of a ruler, at rest in this coordinate system, with one end on A and the other end on D, using our invariant distance rule and treating A and D as infinitesimally separated (we'll learn how to deal with finite separations later). A and D both have the same value of the t coordinate so $dt = 0$. Their x coordinates differ by 10^5 km. So using Eq. 1.2.1 we find the length of this ruler is given by $\sqrt{ds^2} = \sqrt{(10^5\text{km})^2} = 10^5$ km.

Box 1.2.1

Exercise 2.1.1: Use Eq. 1.2.1 to find the time that elapses on a clock that freely falls from D to C. Note that in this coordinate system the clock is not moving at all; i.e., its spatial coordinate is not changing value. Show your work. You should find that you get that the time elapsed is 1 second.

Exercise 2.1.2: Consider the points A and C. Is the separation spacelike or timelike? Evaluate, using Eq. 1.2.1, $\sqrt{-ds^2/c^2}$ if it is timelike and $\sqrt{ds^2}$ if it is spacelike. What does this mean physically in terms of either a clock or a ruler? Answer with a complete sentence.

You should have seen from these two exercises that the time elapsing on a clock traveling from A to C would be *less* than 1 second, whereas a "stationary" clock (one at fixed spatial coordinate) traveling from D to C would show 1 second of time elapsed. This is the phenomenon of time dilation you have studied before, often qualitatively summarized with the statement, "moving clocks run slowly."

It may seem cumbersome to use Eq. 1.2.1 to calculate the time elapsed on a clock that goes from D to C, or the distance between A and D, when the answers seem obvious. However, we will also use spacetimes where the answers are not so obvious, and we really need the guidance of our invariant distance rule. In general, the coordinate values themselves mean nothing -- they are merely labels. Physical meaning becomes apparent only through use of an invariant distance rule. Keep this in mind.

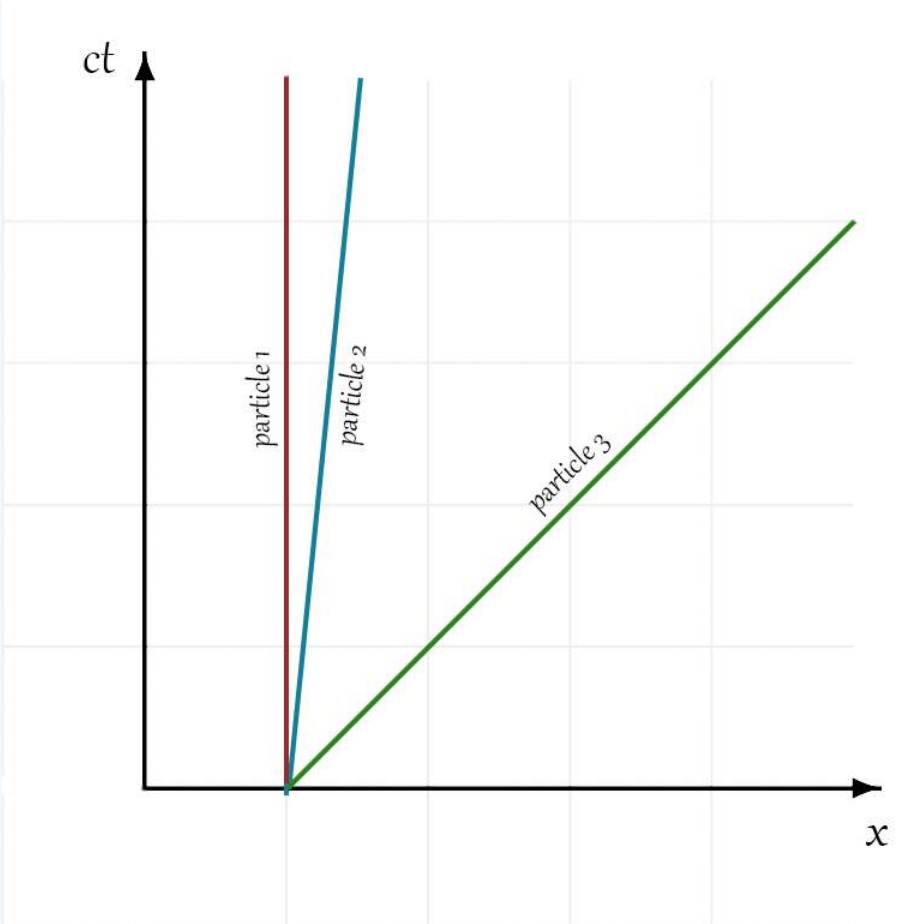
The physical interpretation of ds^2 may seem a bit complicated. But I think it is stated as simply as it can be. The statement for spacelike separations looks particulary cumbersome. Why do we need to specify something about the speed of the ruler? You might recall that there is a phenomenon called "Lorentz contraction;" i.e., that the length of an object in the direction of its motion depends on its speed. So we do need to specify how the ruler is moving -- and here we do so by specifying the frame in which it is at rest.

So far we have used nothing of the general theory of relativity -- the above is all from the special theory of relativity which I assume you have already studied. It might be stated in a form you are unfamiliar with. For example we have emphasized the invariant distance and said nothing about Lorentz transformations, and we have restricted ourselves to infinitesimal, rather than finite, invariant distances. The manner of presentation is intentional, as it makes it easy to extend beyond the framework of special relativity, to that of Einstein's general theory.

Box 1.2.2

Exercise 2.2.1: For the spacetime specified by Equation 1.2.1. On a plot of x vs. t (what we call a spacetime diagram, like the figure above) draw the trajectory of a particle that is not moving, one that is moving slowly, and then of one that is moving at the speed of light. Place the x-coordinate on the horizontal axis, as is the usual convention.

Answer



We cheated a bit here and made a x vs. ct plot so that a particle moving at the speed of light has a slope of 1.

Principles of the General Theory of Relativity

The general theory of relativity (GR) is a theory of gravitation that is conceptually quite different from the Newtonian theory of gravitation. In the Newtonian theory the motions of objects through space and time are understood as being due to a gravitational force that accelerates them. In GR these motions are understood as resulting from an altered geometry of spacetime, with that geometry determined by the matter/energy distribution.

In the general theory of relativity, it is not in general true that one can label space and time so that the invariant distance rule is given by Eq. 1.2.1, or its generalization to multiple spatial dimensions. In fact, the presence of mass distorts the spacetime so that there is no labeling that will make Eq. 1.2.1 true everywhere. However, it is still true in the general theory, like in the special theory, that one can label the spacetime with coordinates, and that their physical meaning is given by a rule specifying the square of the invariant distance between any two infinitesimally-separated points. Further, ds^2 has the same physical interpretation in the general theory as written above.

What do we mean by the geometry of spacetime? That geometry is specified by a rule for the invariant distance that is a function taking in any pair of infinitesimally-separated points and outputting the square of the invariant distance, ds^2. We see above an example of this rule for the special case of a 1+1-dimensional Minkowski spacetime.

How does this geometry dictate how matter moves? For the most part we will only be concerned with the motion of light, which, as we discuss in the next chapter, follows trajectories with $ds^2 = 0$. More generally, a particularly succinct way of stating the rules of force-free motion is as follows: an object that freely falls from event A to event B does so along a spacetime trajectory that extremizes the time elapsed on a clock traveling with the object. How to convert an extremum principle such as this to equations of motion is presented in the optional chapter 1.5. Presumably you have seen this before in the case of the action principle of classical

mechanics and the resulting Euler-Lagrange equations. We also intend to create an additional optional chapter with opportunities to practice use of this extremized-time principle in various spacetimes.

How do matter and energy influence the spacetime geometry? The Einstein Field Equations answer this question; they are beyond the scope of this course.

The Simplest Expanding Spacetime

The invariant distance rule above (Equation 1.2.1) is for a *static* spacetime. It is appropriate for a spacetime in which there is no matter or energy. If instead we have a spacetime with mass uniformly distributed throughout it, then the geometry would be different; i.e., one could no longer label every point in the spacetime with a value of t and of x such that Equation 1.2.1 was true. One could however label the spacetime with t and x such that this is true:

$$ds^2 = -c^2dt^2 + a^2(t)dx^2 \tag{1.2.2}$$

with $a(t)$ a function of time. We call $a(t)$ the "scale factor." The exact behavior of $a(t)$ depends, as we will see, on both initial conditions and the properties of the material supplying that uniform distribution of mass. If $\dot{a} > 0$ the universe is expanding. If $\dot{a} < 0$ it is contracting.

This extremely simple model of a spacetime is one with a high degree of relevance to our own. Of course we have three spatial dimensions instead of one, but for calculating some important observables that difference is irrelevant. And also our universe is clumpy; i.e., the distribution of mass is not spatially uniform. However, as emphasized in the Overview chapter, our universe does appear to be highly uniform on large scales and at early times. This model is well worth studying. Let's do that with the following exercises and into the next chapter.

Box 1.2.3

Exercise 2.3.1: Imagine a very small ruler instantaneously at rest in the x, t coordinate system of Equation 1.2.2 at time $t = t_1$, with one end at location $x = x_1$ and its other end at $x = x_1 + dx_1$. How long is the ruler?

Answer

"at time $t = t_1$" so $dt = 0$, so $ds = a(t)dx$. Since the ruler is at rest in the given coordinate system its length is indeed given by ds at time t_1. Therefore the length of the ruler is $ds = a(t_1)dx_1$.

Box 1.2.4

Exercise 2.4.1: How much time elapses on a clock on a trajectory of constant x, from $t = t_1$ to $t = t_2$ for a spacetime and coordinate system with invariant distances given by Equation 1.2.2? Hint: you have only been given a rule for infinitesimal separations and this is a finite one. However, you can break up the finite interval into infinitesimal ones and then integrate.

Answer

"constant x" so $dx = 0$, and then $ds = cdt$. Therefore the time elapsed on the clock is

$$\int \frac{1}{c}\sqrt{-ds^2} = \int_{t_1}^{t_2} dt = t_2 - t_1.$$

Box 1.2.5

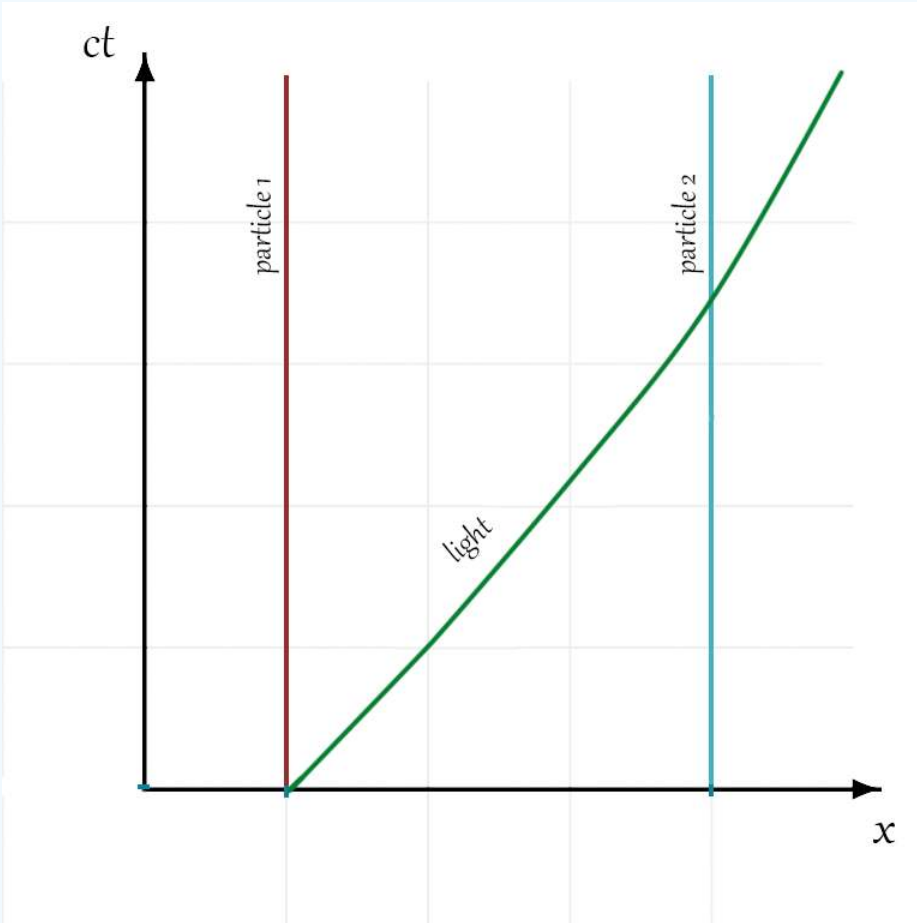
Exercise 2.5.1: Still assuming Equation 1.2.2, draw the paths through spacetime of a pair of particles that are separated from each other and that are not "moving" -- that is, their x coordinate values are not changing over time. Assume $a(t)$ is an increasing function of time. What do you notice about the distance between them and how it evolves over time? Be careful not to confuse "distance between them" with the difference in the values of their spatial coordinates.

Answer

No solution available yet

Exercise 2.5.2: Now, add in the trajectory of a light ray passing from one of these particles to the other. While sketching it out, remember that $a(t)dx$ is the distance traversed (as measured by an observer at rest in the x, t coordinate system) as the time coordinate changes by dt, which is the time elapsed as measured by an observer at rest in the x, t coordinate system. In this x vs. t diagram, does light travel in a straight line?

Answer



Since the amount of distance corresponding to a given dx is changing with time, the slope of the photon's world line is changing with time.

You should have seen in the box above that light does not travel on a straight line in this expanding spacetime as labeled with the t, x coordinates. This is kind of annoying, and something we will address in the next chapter.

An interesting question to ask about an expanding spacetime is whether the universe ever had, in the past, the scale factor equal to zero. If it did, then at this time all pairs of points in space would have zero separation between them -- quite an extreme situation. Just to get some practice working things out in an expanding spacetime, practice that will be useful later, let's assume $\dot{a} = \kappa/a$ for κ some positive constant and see if such a universe ever had $a = 0$. Let us call the time since $a = 0$, Δt. We can then write

$$\Delta t = \int dt = \int_0^{a(t)} da/\dot{a} = \int_0^{a(t)} da(a/\kappa) = a^2(t)/(2\kappa). \tag{1.2.3}$$

Since the integral converged, we find that with the assumption given, namely $\dot{a} \propto 1/a$, the answer is yes, a finite time in the past the scale factor had the value 0. This is the singularity of the big bang. In such spacetimes we usually choose to call the zero point of time ($t = 0$), the time when $a = 0$. [Note that this Δt is the time that would elapse on a stationary clock; i.e., a clock with a fixed spatial coordinate.]

Also note that we made progress with this calculation by replacing dt with $dt = da/\dot{a}$. This is a trick we will use many times to calculate a variety of things.

Another question we can ask is, "how far has light traveled since the beginning." It's interesting because nothing travels faster than the speed of light, so this tells us what the maximum distance is that any signal can propagate. We call this distance the "past horizon." Let's once again assume, for definiteness, $\dot{a} = \kappa/a$ and calculate how far light can travel. We know that for light $ds^2 = 0$ so we have $c^2dt^2 = a^2(t)dx^2$ and therefore $cdt/a(t) = dx$ so we can write

$$\Delta x = \int dx = \int cdt/a = c\int_0^{a(t)} da/(a\dot{a}) = \frac{c}{\kappa}\int_0^{a(t)} da = \frac{c}{\kappa}a(t) \tag{1.2.4}$$

(where you'll note we used the same trick again to convert an integral over time to an integral over the scale factor). Therefore we know the coordinate distance that light has traveled, Δx. That coordinate distance corresponds to a physical distance, at time t, of $a(t)\Delta x = \frac{c}{\kappa}a^2(t)$.

HOMEWORK Problems

Problem 1.2.1

Derive the phenomenon of Lorentz contraction using the invariance of the invariant distance. [Do not assume an expanding universe; assume $ds^2 = -c^2dt^2 + dx^2$]. The trick to doing this is careful choice of the two events (points in spacetime) for which to calculate their invariant distance. Imagine a ruler moving with respect to an observer at speed v, with the ruler oriented so that it is parallel to the relative velocity. Take event 1 to be when/where the front end of the ruler is at the same spacetime location as the observer, and event 2 to be when/where the back end of the ruler is at the same spacetime location as the observer. By calculating the invariant distance in the observer's rest frame and the ruler's rest frame you should find that the length of the ruler as determined by the observer is $L' = L/\gamma$ where L is the length of the ruler in its rest frame.

Problem 1.2.2

Assume that the scale factore evolves via $\dot{a} = \kappa a$ for κ a positive constant. (Note that this is a *different* assumption than the previous $\dot{a} = \kappa/a$). Show that in this spacetime the universe never has $a = 0$. Do so by showing that the amount of time between $a = 0$ and any finite a is infinite; i.e., show that the appropriate definite integral does not converge.

Problem 1.2.3

Assume $ds^2 = -c^2dt^2 + a^2(t)dx^2$ and once again that $\dot{a} = \kappa a$ for κ a positive constant. Our universe appears to be moving asymptotically toward such a case (although except with a 3-dimensional space instead of a 1-dimensional space). Determine what we call the "future horizon." If a light signal is sent out at time t_1 from x_1, in the positive x direction, to what value of x_2 will it get given an infinite amount of time? The distance between x_1 and x_2 at time t_1, $a(t_1)(x_2 - x_1)$, is called the future horizon.

This page titled 1.2: Spacetime Geometry is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

- **1.36: The Simplest Expanding Spacetime** by Lloyd Knox is licensed CC BY 4.0.
- **2.5: S05. Spacetime - SOLUTIONS** by Lloyd Knox is licensed CC BY 4.0.

1.3: Redshifts

We introduced an expanding spacetime in the previous chapter. Now we begin to work out observational consequences of living in such a spacetime. In this and the next few chapters we will derive Hubble's Law, $v = H_0 d$. In fact, we will derive a more general version of it valid for arbitrarily large distances.

Here we continue to work with just one spatial dimension. Assuming a homogeneous universe we have the same invariant distance rule as before:

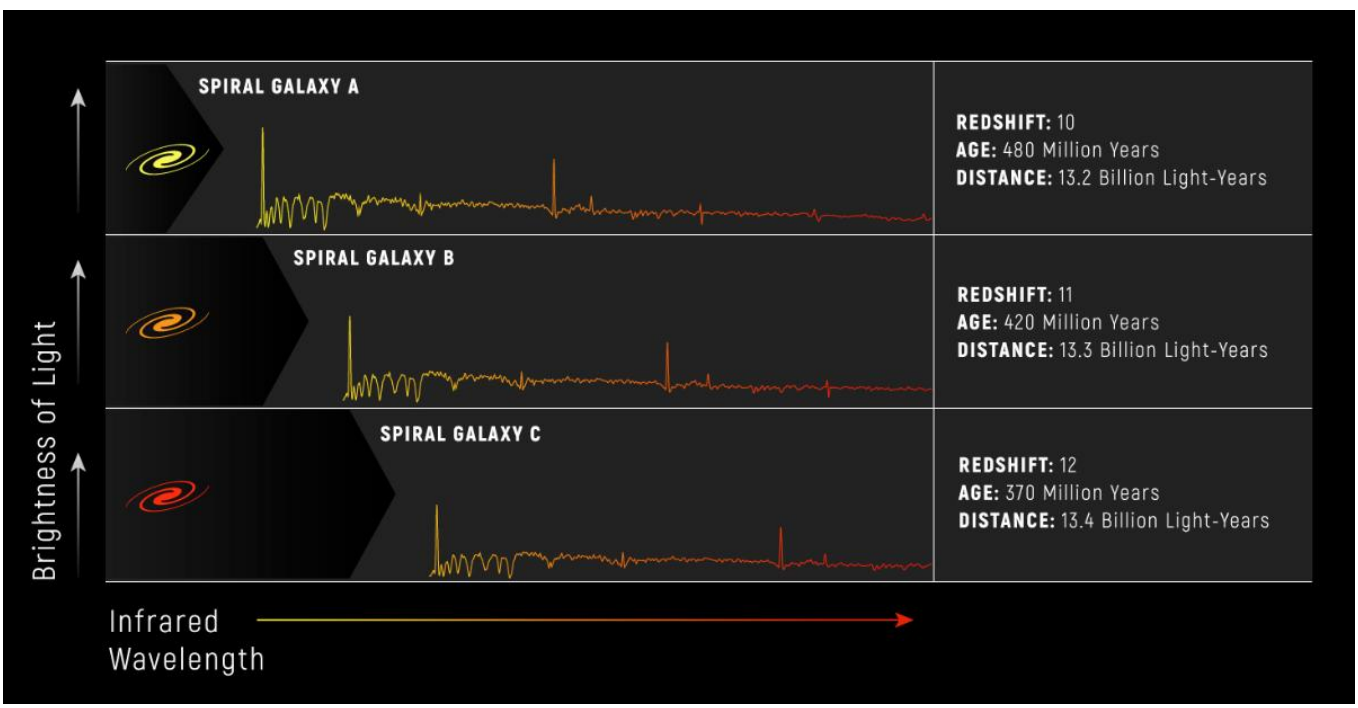
$$ds^2 = -c^2 dt^2 + a^2(t) dx^2. \tag{1.3.1}$$

We derive in this chapter how the wavelength of light stretches over time, due to the expansion of the universe. This is called the cosmological redshift. We will find that the cosmological redshift is a very simple function of the scale factor at the time of emission divided by the scale factor at the time of reception. In other words, redshift tells us how much the universe expanded since light left the object we are observing.

In general, whether its origin is the expansion of space or the motion of an object, redshift is a quantification of the stretching of the wavelength of light. We use the symbol z to denote redshift and define it as:

$$z \equiv (\lambda_{\text{received}} - \lambda_{\text{emitted}})/\lambda_{\text{emitted}}. \tag{1.3.2}$$

We see in the figure below (from NASA and Space Telescope Science Institute) some model spectra of galaxies with redshifts of 10, 11, and 12. One can see in the image that the same spectral features appear at longer wavelengths as the redshift increases. These features all had the same wavelength when emitted, but those wavelengths get stretched by the expansion to longer wavelengths. The further away the objects, the greater the stretching that occurs. We will explore this relationship between distance and redshift in chapter 1.7



Next we want to derive the relationship between redshift and expansion for light that leaves an object at rest at time t_e and is observed today by an observer at rest at time t_r. To help us get there we will first introduce what we call "conformal time" and then discuss the value of the invariant distance on paths traveled by light.

Note

By "at rest" we mean at rest in the coordinate system that has an invariant distance of the form in Equation 1.3.1. This frame is called "cosmic rest." Note that for a frame that is moving with respect to cosmic rest, slices of constant time would be different, and we'd no longer have this simple form where the scale factor only depends on the time coordinate.

Conformal Time

You will recall from the previous chapter that in an expanding spacetime, light does not travel on straight lines. For calculational purposes, as we will see, it would be nicer if it did move on straight lines. Well, with a change of coordinates, it can! For this purpose, we introduce a coordinate called "conformal time." The conformal time, η, is defined via $d\eta = dt/a(t)$. In a conformal time diagram, for the expanding spacetime with which we have been working, light trajectories are straight lines. We will find that this is a very useful property.

Box 1.3.1

Exercise 3.1.1: Given a spacetime described by Equation 1.3.1, work out the invariant distance specified for η, x labeling instead of t, x labeling. You should find $ds^2 = a^2(\eta)[-c^2 d\eta^2 + dx^2]$ where by $a(\eta)$ we just mean $a(t(\eta))$. Note that $a(\eta)$ is simply shorthand for $a(t(\eta))$.

Answer

Substituting in $dt = a(t)d\eta$ to Equation 5.2 and factoring out $a^2(t)$ gives us

$$ds^2 = -c^2 a^2(t)[d\eta^2 + dx^2].$$

Assuming a one-to-one correspondence between t and η (which one would have in an expanding universe given definition of $d\eta$) we can use $a(\eta) \equiv a(t(\eta))$ in its place and write

$$ds^2 = a^2(\eta)[-c^2 d\eta^2 + dx^2]$$

Beware: in previous versions of this textbook we used τ for conformal time and it's possible that in some places we still have the old notation. I found that the use of τ caused some confusion for students because they have seen that used before for proper time, and these are not the same thing.

The Invariant Distance for Light Trajectories

The invariant distance for a path taken by a photon is always zero. You know this is true in special relativity where $ds^2 = -c^2 dt^2 + dx^2$. Light always travels at the speed of light, so we always have $dx = \pm c dt$, which in turn ensures that $ds^2 = -c^2 dt^2 + c^2 dt^2 = 0$. To see that this is true more generally, first note that for an observer at rest in a given coordinate system, and given our physical interpretation of the invariant distance, the equation for the invariant distance can always be written schematically in terms of the trajectory of a moving object as

$$ds^2 = -c^2(\text{infinitesimal time elapsed})^2 + (\text{infinitesimal spatial distance traversed})^2$$

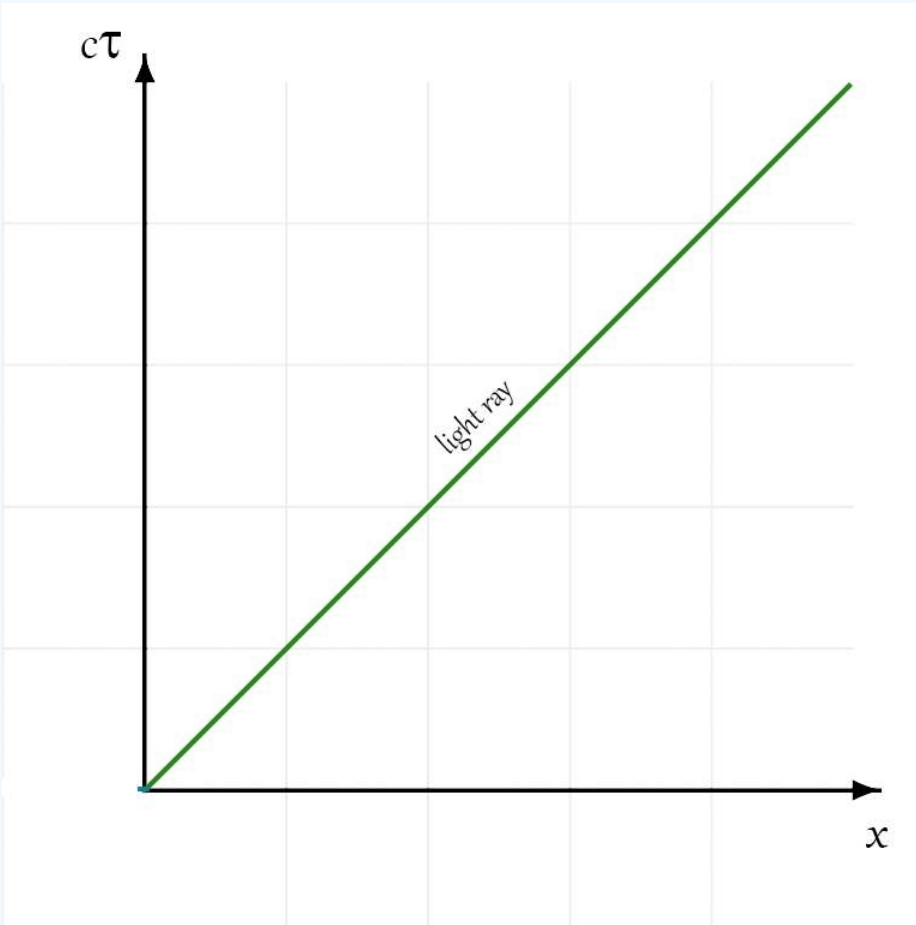
where by "infinitesimal time elapsed" we mean as measured by a clock that is *not* moving in the given coordinate system and by "infinitesimal spatial distance traveled" we mean as measured by a ruler that is not moving in the given coordinate system. You can immediately see that this is true in two steps. First, consider a purely spatial separation (in the given coordinate system). The invariant distance rule then tells you that the invariant distance squared is the square of the distance between the two points as measured by a ruler at rest in that coordinate system. This verifies the second term on the right-hand side of the above equation. Second, consider a purely temporal separation. The invariant distance rule tells you that the square of the invariant distance divided by $-c^2$ gives the square of the time elapsed on a clock moving between these two points. This verifies the first term on the right-hand side of the above equation.

You know that in the special theory of relativity that all observers see light moving at speed c. A version of this is true in general relativity. We just have to be careful to state it more locally. In the general theory, any observer seeing light pass right by them (through their location), will see it traveling with speed c. Since this is true for all observers, it is also true for an observer at rest in the given coordinate system, so they will see (infinitesimal spacial distance traversed) = $c\times$ (infinitesimal time elapsed). Putting this together with the above schematic equation for ds^2 we see that for a light trajectory, $ds^2 = 0$.

Box 1.3.2

Exercise 3.2.1: Draw how light rays move on a plot of x vs. η assuming our homogeneous expanding spacetime with one spatial dimension. Start from $ds^2 = 0$ to find the relationship between $d\eta$ and dx, then draw a trajectory consistent with that relationship.

Answer



The theoretical relationship between redshift and scale factor history

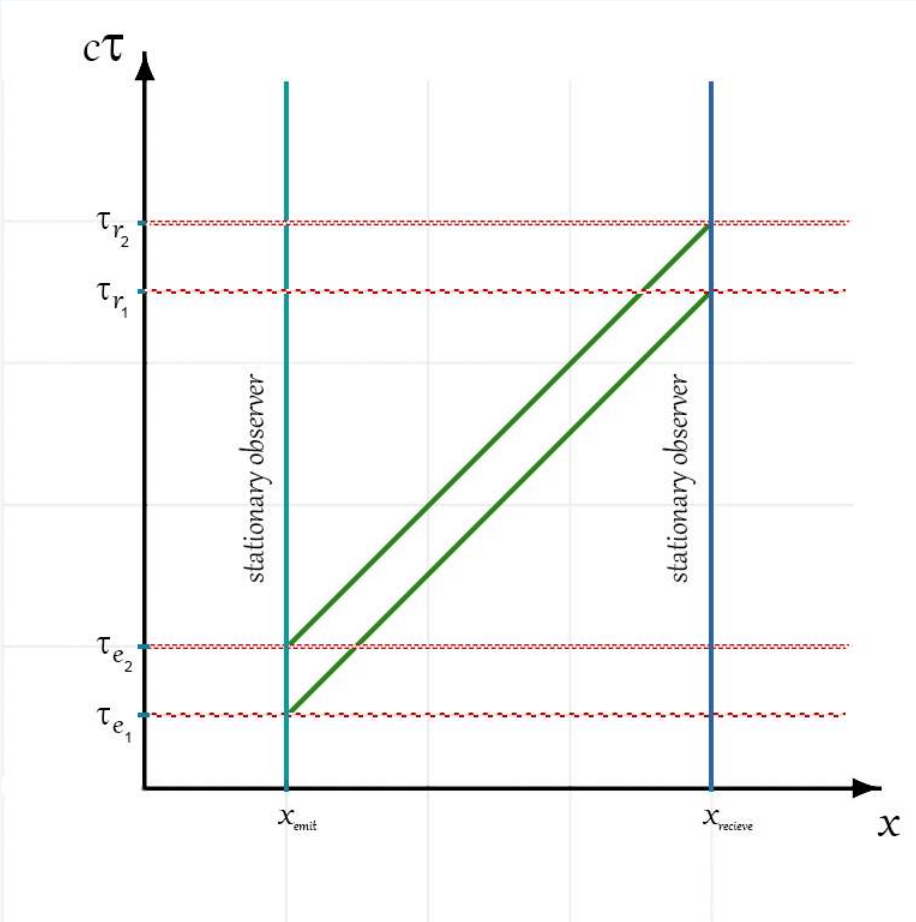
We are all set up now to show, through the next three exercises, how redshift is related to the expansion that occurs between the time of emission of the light $t_{\rm e}$, and the time of reception of the light, $t_{\rm r}$. We will find the following very simple relationship stating that the wavelength has been increased by exactly the same factor by which the scale factor has increased; i.e.,

$$1 + z \equiv \lambda_{\rm r}/\lambda_{\rm e} = a(t_{\rm r})/a(t_{\rm e}). \tag{1.3.3}$$

Box 1.3.3

Exercise 3.3.1: Show that if Δt_e is the time interval between emitted light pulses (as measured by a stationary observer located where the emission is happening) and if Δt_r is the time interval between reception of first pulse and second pulse (as measured by a stationary observer located where the reception is occurring) then $\Delta t_r/\Delta t_e = a(t_r)/a(t_e)$. To do so, use conformal time defined by $d\eta = dt/a(t)$ and draw the pulse trajectories on an x vs. η diagram. By stationary observer we mean one that is at a constant value of x; i.e., is at rest in the cosmic rest frame. You can assume that Δt_r and Δt_e are very short time scales compared to the time scale over which $a(t)$ changes appreciably. Practical applications of this result often have $a(t)$ changing on billion-year time scales and the Δts shorter than nanoseconds so this assumption is well justified!

Answer



It's clear from the graphic that $\Delta\tau_e = \Delta\tau_r$. We can integrate up $d\tau = dt/a(t)$ by approximating $a(t)$ as constant over these short time intervals to get $\Delta\tau = \Delta t/a(t)$. Then we can easily see that $\Delta t_r/\Delta t_e = a(t_r)/a(t_e)$.

Box 1.3.4

Exercise 3.4.1: Imagine propagation of electromagnetic waves. Use the above result to show that the wavelength of the waves emitted at time t_r and observed at time t_e is stretched so that:

$$z \equiv (\lambda_{\text{received}} - \lambda_{\text{emitted}})/\lambda_{\text{emitted}} = a(t_r)/a(t_e) - 1$$

Hint: think about what happens to the period of the wave first, and then go from that to wavelength using the fact that wavelength is proportional to period.

Answer

Imagine successive crests of a single wave. Map one crest, and then the subsequent one, separated in time by the period of the emitted wave T_e, on to the two pulses you considered in the previous exercise. The time between the pulses upon emission is $\Delta t_e = T_e$. The time between the reception of the first pulse and reception of the second pulse is the period of the wave upon reception $T_r = \Delta t_r = (a(t_r)/a(t_e))\Delta t_e = (a(t_r)/a(t_e))T_e$. Wavelength is proportional to period so we also have $\lambda_r = (a(t_r)/a(t_e))\lambda_e$ or

$$\frac{\lambda_r}{\lambda_e} = \frac{a(t_r)}{a(t_e)}.$$

We already know that $z \equiv (\lambda_{\text{received}} - \lambda_{\text{emitted}})/\lambda_{\text{emitted}}$, so we can rewrite it as

$$z \equiv \frac{\lambda_r}{\lambda_e} - 1 = \frac{a(t_r)}{a(t_e)} - 1$$

Box 1.3.5

Exercise 3.5.1: The most distant object for which a redshift has been measured is called a gamma-ray burst and it has a redshift of $z = 8.2$. By what factor has the universe expanded since light left that object?

Answer

The universe has expanded by a factor

$$a(t_r)/a(t_e) = 1 + z = 1 + 8.2 = 9.2.$$

We have just worked out the amazing fact that if we can identify a spectral line and measure its wavelength, then we can directly determine how much the universe has expanded since light left the object. Rearranging the result from Exercise 2.2.1 a little we can express this amazing fact as an equation:

$$1 + z = \lambda_r/\lambda_e = a(t_r)/a(t_e). \tag{1.3.4}$$

Look again at the figure above and notice that the more distant the object, the higher the redshift. This is because the more distant the object, the more time it took light to get to us from the object, and therefore the smaller the scale factor was at the time the light left the object, t_e.

Before closing this section, we remind the reader of how redshifts due to the Dopper effect are related to speed -- a result we will use later in deriving Hubble's Law. If a source is moving away from you at speed v and emitting pulses with period T, then the second pulse has to travel a distance vT further to get to you than was the case for the first pulse. So it's arrival will be delayed by a time vT/c. Thus the period for the arriving pulses is $T(1 + v/c)$. Since wavelength is proportional to period this means the wavelength is stretched by a factor of $1 + v/c$, which means, by definition of the redshift z, that $z = v/c$. Note that our derivation has ignored the effect of relativistic time dilation. If the source had period T at rest, then if it were moving with speed v with respect to us, in our frame the period would be stretched to γT and so the complete expression for z from the Doppler effect is $1 + z = \gamma(1 + v/c)$. But we are only interested in this expression for small v/c, for which $\gamma \sim 1$.

Summary

1. In an expanding homogeneous and isotropic universe, the ratio between the wavelength of light emitted by an observer at cosmic rest, λ_e at time t_e and the wavelength as measured by an observer at cosmic rest λ_r at time t_r is given by

$$1 + z \equiv \lambda_r/\lambda_e = a(t_r)/a(t_e) \tag{1.3.5}$$

   where $a(t)$ is the scale factor at time t and the above equation defines the redshift z. We can think of this as the wavelength stretching with the expansion.
2. In a non-expanding universe, a source moving away from an observer with $v/c \ll 1$ has its light redshifted (wavelength stretched) by

$$\lambda_r/\lambda_e = 1 + v/c \tag{1.3.6}$$

   which is the normal Doppler effect you have studied before. What this implies is that if we interpret a small redshift z caused by the expansion of space as due to an ordinary motion-induced Dopper effect we will set $z = v/c$. We will use this relationship later in our derivation of Hubble's Law: $v = H_0 d$. The next thing we need for Hubble's law is how to measure distances, and how those measurements depend on theoretical quantities.

Additional Resources

You can look at images and spectra of galaxies in the Sloan Digital Sky Survey here. The spectra include identification of emission and absorption lines (with the atoms or ions responsible for them) and measurements of the redshift. To find the spectra and images, look for the table and click on the Object ID.

Quasar absorption line systems have particularly interesting spectra. You can read about them here.

HOMEWORK Problems

Problem 1.3.1

A phenomenon closely related to cosmological redshift is cosmological time delay. Light signals arriving from events at the same location, but separated in time so one happens at t_e and the other at $t_e + \delta t_e$, are separated by a larger time interval, $\delta t_r > \delta t_e$. Derive how $\delta t_r / \delta t_e$ depends on $a(t_r)/a(t_e)$. Use a conformal time diagram as part of your derivation. You can assume δt_r and δt_e are both much smaller than time scales over which $a(t)$ changes appreciably.

Problem 1.3.2

A supernova is observed today to take 20 days to reach peak brightness from the beginning of the explosion. It has a redshift of $z = 1$. At the location of the supernova, back when the explosion occurred, how many days did it take to go from beginning of explosion to peak brightness?

This page titled 1.3: Redshifts is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

- **1.36: The Simplest Expanding Spacetime** by Lloyd Knox is licensed CC BY 4.0.
- **1.37: Redshifts** by Lloyd Knox is licensed CC BY 4.0.
- **2.5: S05. Spacetime - SOLUTIONS** by Lloyd Knox is licensed CC BY 4.0.
- **2.6: S06. Redshifts - SOLUTIONS** by Lloyd Knox is licensed CC BY 4.0.

1.4: Spatially Homogeneous and Isotropic Spacetimes

The Cosmological Principle

The Einstein field equations are extremely difficult to solve in generality. The first attempts at solving these equations for the universe as a whole thus involved extreme idealization. In the immediate years after Einstein presented his theory of general realtivity, several people used what you might call the most spherical cow approximation of all time: they approximated the whole universe as completely homogeneous; i.e., absolutely the same everywhere. The 'cosmological principle' is simply the assertion that the universe is homogeneous (invariant under translations) and isotropic (invariant under rotation). Model spacetimes with this high degree of symmetry are still of interest to us today because, as discussed in the Overview, on large scales and at early times, the universe is in fact very close to being homogeneous and isotropic.

Three-dimensional Homogeneous and Isotropic Spaces

So far we have worked with spacetimes with just one spatial dimension. But to get to Hubble's law we need to know how to measure distances. And for our particular method of measuring distances, we are going to need to work with more than one spatial dimension, as we will explain later. So the time has come to think about additional spatial dimensions. There appear to be three spatial dimensions, so let's start there.

In Minkowski space, the square of the invariant distance, ds^2, between spacetime point (t, x, y, z) and another one at $(t + dt, x + dx, y + dy, z + dz)$ is given by:

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2. \tag{1.4.1}$$

In spherical coordinates the above expression for the invariant distance becomes:

$$ds^2 = -c^2dt^2 + dr^2 + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right). \tag{1.4.2}$$

The above is a special case valid for static (non-expanding) spacetimes with Euclidean spatial geometries. The invariant distance for *any* spatially homogeneous and isotropic Universe can be written as:

$$ds^2 = -c^2dt^2 + a^2(t)\left[\frac{dr^2}{1-kr^2} + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right)\right]. \tag{1.4.3}$$

Such spacetimes are known as Friedmann-Robertson-Walker (FRW) models, or sometimes just Robertson-Walker, and sometimes Friedmann-Robertson-Walker-Lemaitre models.

Derivation

We can construct the homogeneous and isotropic three-dimensional space and derive its invariant distance rule, at least for the case of $k > 0$, by embedding it in a 4-dimensional Euclidean space. In a 4-dimensional Euclidean space we can have a coordinate system consisting of three dimensions x, y, z, that are all orthogonal to each other, and a fourth we will call w that is orthogonal to each of the x, y, and z directions. Impossible as this is to visualize, we can describe it mathematically. The distance between w, x, y, z and $w + dw, x + dx, y + dy, z + dz$ is given by

$$d\ell^2 = dw^2 + dx^2 + dy^2 + dz^2. \tag{1.4.4}$$

In this 4-dimensional space, we construct a three-dimensional subspace that is the set of points all the same distance, R, from a common center. Let's center it on the origin so our subspace satisfies this constraint:

$$w^2 + x^2 + y^2 + z^2 = R^2. \tag{1.4.5}$$

This subspace is homogeneous (all points are the same) and isotropic (all directions are the same). You can see that this is true by imagining it's two-dimensional analog, a sphere, which is the set of all points satisfying $x^2 + y^2 + z^2 = R^2$.

It will be helpful at this point to swap out the Cartesian x, y, z for the spherical coordinate system r, θ, ϕ so we have

$$d\ell^2 = dw^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2) \tag{1.4.6}$$

and our constraint equation can be written as

$$w^2 + r^2 = R^2. \tag{1.4.7}$$

From this new version of the constraint equation, we can see that if r changes by some amount then we will necessarily have to have a change in w in order to continue to satisfy the constraint. The exact relationship between differential changes you can easily work out to be $2wdw + 2rdr = 0$ (because changing r by dr ends up changing r^2 by $2rdr$ and likewise for w and dw and since R is fixed $dR^2 = 0$). Using this relationship to eliminate dw^2 from our invariant distance expression, and using the constraint equation to eliminate w^2 in favor of r^2 and R^2 we get

$$d\ell^2 = \frac{dr^2}{1 - r^2/R^2} + r^2(d\theta^2 + \sin^2\theta d\phi^2). \tag{1.4.8}$$

We see that our subspace has an invariant distance expression of the form we were intending to derive, and it is exactly the one introduced above if we make the identification $R^2 = 1/k$.

HOMEWORK Problems

Problem 1.4.1

Re-do the above derivation leading to $d\ell^2 = dw^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2)$ but for a 2-dimensional space instead of a 3-dimensional space. Instead of spherical coordinates (r, θ, ϕ), use cylindrical coordinates (r, ϕ).

This page titled 1.4: Spatially Homogeneous and Isotropic Spacetimes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

- **1.37: Redshifts** by Lloyd Knox is licensed CC BY 4.0.
- **1.33: Curvature** by Lloyd Knox is licensed CC BY 4.0.
- **1.36: The Simplest Expanding Spacetime** by Lloyd Knox is licensed CC BY 4.0.

1.5: Euclidean Geometry

This is an optional chapter for those who want to improve their geometry skills. We intend to add one more optional chapter on geometry to provide the opportunity to practice with non-Euclidean spaces -- both the spatial part of FRW and spatial part of Schwarzschild.

The chapter is entirely focused on the Euclidean geometry that is familiar to you, but reviewed in a language that may be unfamiliar. The new language will help us journey into the foreign territory of Riemannian geometry where space is curved. Our exploration of that territory will then help you to drop your pre-conceived notions about space and to begin to understand the broader possibilities -- possibilities that are not only mathematically beautiful, but that appear to be realized in the natural world. Perhaps of particular note is our discussion of a result of the calculus of variations, the connection between an insistence that a path extremize some integral property, and the differential equations governing the path that necessarily follow from that insistence. Such an extremum principle comes up in classical mechanics, where the resulting equations are the Euler-Lagrange equations. General relativity also has such an extremum principle for force-free motion, a principle we use in chapter 12.

According to Euclidean geometry, it is possible to label all space with coordinates *x*, *y*, and *z* such that the square of the distance between a point labeled by x_1, y_1, z_1 and a point labeled by x_2, y_2, z_2 is given by $(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2$. If points 1 and 2 are only infinitesimally separated, and we call the square of the distance between them $d\ell^2$, then we could write this rule, that gives the square of the distance as

$$d\ell^2 = dx^2 + dy^2 + dz^2 \tag{1.5.1}$$

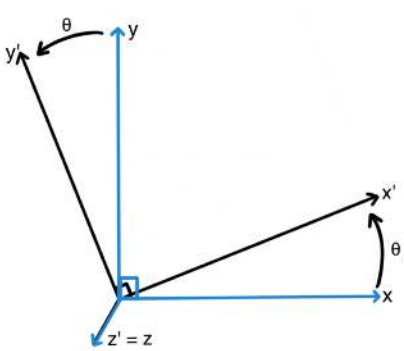
This rule has physical significance. The physical content is that if you place a ruler between these two points, and it is a good ruler, it will show a length of $d\ell = +\sqrt{d\ell^2}$. Since it is difficult to find rulers good at measuring infinitesimal lengths, we can turn this into a macroscopic rule. Imagine a string following a path parameterized by λ, from $\lambda = 0$ to $\lambda = 1$, then the length of the string is $\int_0^1 d\lambda(d\ell/d\lambda)$. That is, every infinitesimal increment $d\lambda$ corresponds to some length $d\ell$. If we add them all up, that's the length of the string.

Box 1.5.1

Exercise 5.1.1: Find the distance along a path from the origin to (x,y,z) = (1,1,1) where the path is given by

$$x(\lambda) = \lambda, y(\lambda) = \lambda, z(\lambda) = \lambda \tag{1.5.2}$$

There are many ways to label the same set of points in space. For example, we could rotate our coordinate system about the *z* axis by angle θ (with positive θ taken to be in the counterclockwise direction as viewed looking down toward the origin from positive z) to form a primed coordinate system with this transformation rule:



$$z' = z \tag{1.5.3}$$

$$y' = -x\sin\theta + y\cos\theta \tag{1.5.4}$$

$$x' = x\cos\theta + y\sin\theta \tag{1.5.5}$$

Under such a re-labeling, the distance between points 1 and 2 is unchanged. Physically, this has to be the case. All we've done is used a different labeling system. That can't affect what a ruler would tell us about the distance between any pair of points. Further, for this particular transformation, the equation that gives us the distance between infinitesimally separated points has the same form.

Figure 1: A counterclockwise rotation of the coordinate system about the z axis by θ creates a new coordinate system which we've labeled with primes. The z axis comes out of the screen and is identical to the z' axis. As is true for any point in space, point 1 can be described in either coordinate system, by specifying (x_1, y_1, z_1) or (x_1', y_1', z_1') with the relationship between the two given by the equations to the right.

Box 1.5.2

Exercise 5.2.1: Show that the distance rule of Equation 1.5.1 applied to the prime coordinates,

$$(d\ell')^2 = (dx')^2 + (dy')^2 + (dz')^2$$

gives the same distance; i.e, show that $d\ell' = d\ell$. [Warning: θ is not a coordinate here. It specifies the relationship between the coordinate systems. So, e.g., $dx' = dx\cos\theta + dy\sin\theta$.] Because this distance is invariant under rotations of the coordinate system, we call it the invariant distance.

We want to emphasize that the labels themselves, *x*, *y*, *z* or *x'*, *y'*, *z'* have no physical meaning. All physical meaning associated with the coordinates comes from an equation that tells us how to calculate distances along paths. To drive this point home, note that we could also label space with a value of *x*, *y*, *z* at every point, but do it in such a way that we would have the distance between *x*, *y*, *z* and $x + dx, y + dy, z + dz$ have a square given by

$$d\ell^2 = dx^2 + x^2\left(dy^2 + \sin^2 y dz^2\right) \tag{1.5.6}$$

For many readers, this result would look more familiar if we renamed the coordinates $r = x$, $\theta = y$, and $\phi = z$ so that we get another expression for the invariant distance,

$$d\ell^2 = dr^2 + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right) \tag{1.5.7}$$

This is the usual spherical coordinate labeling of a 3-dimensional Euclidean space by distance from origin, *r*, a latitude-like angle, θ, and a longitudinal angle, ϕ. The transformation between the two coordinate systems is given by

$$z = r\cos\theta \tag{1.5.8}$$

$$y = r\sin\theta\sin\phi \tag{1.5.9}$$

$$x = r\sin\theta\cos\phi \tag{1.5.10}$$

Box 1.5.3

Exercise 5.3.1: Show that the invariant distance given by the equation $d\ell^2 = dx^2 + dy^2$, the 2-D version of 1.5.1, and the invariant distance given by the equation $d\ell^2 = dr^2 + r^2 d\phi^2$, the 2-D version of 1.5.7, are consistent if the coordinates are related via:

$$x = r\cos\phi$$

$$y = r\sin\phi$$

Hint: use the chain rule, so that, e.g., $dx = dr\cos\phi - r\sin\phi d\phi$. (Note that the coordinate transformation equations here are obtained from the 3-dimensional case by setting $\theta = \pi/2$.)

In preparation for thinking about non-Euclidean spaces, we are going to go through how one could construct a labeling of a two-dimensional Euclidean space in polar coordinates, r, ϕ. Our construction starts with what will look like an unusual way of defining r. We define r based on the circumference of the circle rather than the distance from the origin, for reasons that will become clear later.

First we choose a center to our coordinate system. Then we label all points with r that are equidistant from that center and form a circle with circumference $C = 2\pi r$. Thus to label space with the appropriate value of r, one takes a string, ties one end down at the center, and marks out all the points that can be just reached by the other end of the string, when it is pulled straight. Then one measures the circumference of the resulting circle and labels the points on this circle with a value of r given by $r = C/(2\pi)$. We take strings of varying lengths and repeat again and again to figure out the value of r for every point in the plane.

Next, to label space with ϕ, we take one point on one of the circles and arbitrarily label that one as $\phi = 0$. We pull a string tight from the origin out to this point and beyond, and label all the points along the string with $\phi = 0$. We then march outward from the origin and when we get to a point labeled with radial value r, we make a 90° turn to the left and advance some small distance Δ. We then label this point with $\phi = \Delta/r$. Again we pull a string tight from the origin out to this point and beyond, and label all these points along the string with the same value of ϕ. We then advance another Δ around the circle and repeat, now labeling the nth

iteration with $\phi = n\Delta/r$. In this manner we label all points in the space with values of ϕ. Note that when we have done this $2\pi r/\Delta$ times we will have advanced all the way around the circle (because we will have covered a distance of $2\pi r$) and the change in ϕ will be $(2\pi r/\Delta) \times \Delta/r = 2\pi$.

In a Euclidean space, such a construction leads us to the result (unproven here) that the distance $d\ell$ between two infinitesimally separated points labeled by r, ϕ and $r + dr, \phi + d\phi$ has a square given by

$$d\ell^2 = dr^2 + r^2 d\phi^2. \tag{1.5.11}$$

Note that in our construction we never made any measurement of the distance from the origin to a circle with origin as center and with circumference $C = 2\pi r$. All we know so far is the circumference of the circle. To calculate the distance from the origin to this circle we can apply the above rule for a path that extends from the origin to the circle. Let's say a circle with circumference $C_1 = 2\pi r_1$.

Box 1.5.4

Exercise 5.4.1: Calculate the distance from the origin to the circle with circumference $C_1 = 2\pi r_1$. Do so along a path of constant ϕ using Eq. 1.5.11.

You should have got the unsurprising result that the distance from the origin to the circle with circumference $C_1 = 2\pi r_1$ is r_1. In the next chapter this will get more interesting as we examine a space for which this is *not* the case. We'll see that the distance to a circle with this circumference could be more than r_1 or less than r_1.

We constructed our coordinate system so that as θ goes from 0 to 2π at constant $r = r_1$ a distance is traversed of $2\pi r_1$. Let's now check that our rule for $d\ell$ above, Eq. 1.5.11 is consistent with this construction.

Exercise 5.4.2: Show that the parameterized path $r = r_1, \theta = \lambda$ as λ goes from 0 to 2π covers a distance of $2\pi r_1$ by integrating $d\ell$, as given by Eq. 1.5.11, along this path.

Before going on, we could take a little more care. We have shown that a particular path that takes us from the origin out to $r = r_1$ at constant ϕ has distance r_1. But how do we know this is the shortest path? Here we will demonstrate that there is not a shorter path; the one prescribed is the shortest path possible. To do so, we use a result from the calculus of variations. That result is as follows:

For $J = \int_1^2 d\mu f(q_i, \dot{q}_i, \mu)$ where $\dot{q}_i \equiv dq_i/d\mu$, the path from point 1 to 2 that extremizes J satisfies these equations

$$\frac{d}{d\mu}\left(\frac{\partial f}{\partial \dot{q}_i}\right) = \frac{\partial f}{\partial q_i} \tag{1.5.12}$$

This is a mathematical result with more than one application. In mechanics, the action is given as an integral over the Lagrangian so that

$$S = \int dt L(q_i, \dot{q}_i, t) \tag{1.5.13}$$

with $\dot{q}_i \equiv dq_i/dt$, and because a system passes from point 1 to point 2 along the path that minimizes the action, the path taken will satisfy

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) = \frac{\partial L}{\partial q_i} \tag{1.5.14}$$

which you know as the *Euler-Lagrange equations*.

In the case at hand we have length = $\int d\mu \frac{d\ell}{d\mu}$ where

$$f = \frac{d\ell}{d\mu} = \sqrt{\dot{r}^2 + r^2\dot{\phi}^2} \tag{1.5.15}$$

(note the overdot is differentiation with respect to the independent variable which here is μ again) so the shortest-length path between any two points should satisfy

$$\frac{d}{d\mu}\left(\frac{\partial f}{\partial \dot{r}}\right)=\frac{\partial f}{\partial r} \tag{1.5.16}$$

$$\frac{d}{d\mu}\left(\frac{\partial f}{\partial \dot{\phi}}\right)=\frac{\partial f}{\partial \phi} \tag{1.5.17}$$

These equations are kind of hairy, if you work them out in generality. However, we are testing to see if a particular path satisfies them, the path from the origin to $r=r_1$, and $\phi=\phi_1$ that proceeds at fixed ϕ. We could parameterize this path with $\phi=\phi_1, r(\mu)=\mu r_1$ with μ running from 0 to 1. Note that $\dot{\phi}=0$ which really simplifies the evaluation of the above equations. We will just do one term out of the first equation as an example, and leave evaluation of the rest of the terms as an exercise. In particular, we evaluate $\partial f/\partial r=(r/f)\dot{\phi}^2=0$.

Box 1.5.5

Exercise 5.5.1: Evaluate the three other terms ($\frac{d}{d\mu}\left(\frac{\partial f}{\partial \dot{r}}\right)$, $\frac{d}{d\mu}\left(\frac{\partial f}{\partial \dot{\phi}}\right)$ and $\frac{\partial f}{\partial \phi}$) in the two equations above, and verify that the given path does indeed satisfy these equations, thereby demonstrating that it is the shortest possible path.

Summary

1. Space can be labeled with coordinates. The same space can be labeled with a variety of coordinate systems; e.g., Cartesian or Spherical.
2. The coordinate labelings themselves have no physical meaning. Physical meaning resides in the distances between points, which one can calculate from a rule that relates infinitesimal changes in coordinates to infinitesimal distances.
3. Paths through a space can be parameterized by a single variable; we saw several examples of this.
4. The Euler-Lagrange equations can be used to prove that a particular path is (or is not) one with an extreme value of distance between a pair of points on the path. Usually the extreme is a minimum rather than a maximum.

Homework Problems

Problem 1.5.1

Starting from $d\ell^2=dx^2+dy^2$ prove the Pythagorean theorem that the squares of the lengths of two sides of a right triangle are equal to the square of the hypotenuse. Start off by proving it for a triangle with the right-angle vertex located at the origin, so all three vertices are at $(x,y)=(0,0),(x_1,0)$, and $(0,y_1)$. Be careful to use the distance rule to determine the length of each leg of the triangle, rather than your Euclidean intuition. Let's call the length of the side along the x-axis ℓ_x and similarly the other lengths ℓ_y and ℓ_h. Parameterize each path and perform the appropriate integral over the independent variable you used for the parametrization (like we did with λ in this chapter). Doing so, you should find that $\ell_h^2=\ell_x^2+\ell_y^2$. Having proved the Pythagorean theorem for this specially located and oriented triangle, note that since translations and rotations of the coordinate system leave our invariant distance rule unchanged, you have effectively proved it for all right triangles.

Problem 1.5.2

Prove that the hypotenuse, the straight line from $(x_1,0)$ to $(0,y_1)$ you described in 1.1, is the shortest path between those two points.

Problem 1.5.3

Show that for a primed system that is rotated relative to the unprimed system so that

$$x'=x\cos\theta+y\sin\theta$$

$$y'=-x\sin\theta+y\cos\theta$$

the square of the invariant distance is unchanged; i.e., $dx^2+dy^2=(dx')^2+(dy')^2$.

This page titled 1.5: Euclidean Geometry is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

- **1.32: Euclidean Geometry** by Lloyd Knox is licensed CC BY 4.0.

1.6: Distances as Determined by Standard Candles

For now we move on to the measurement of distances, something we'll also need for the derivation of Hubble's Law and its generalization valid for very large distances. One way to measure a distance is called the "standard candle" method. Assume we have an object with luminosity L where luminosity is the energy per unit time leaving the object. Measuring the flux (energy/unit time/unit area) will give us a way to figure out the distance to the object assuming it is emitting isotropically. The further away it is, the weaker the flux will be.

To determine the relationship between luminosity, flux and distance we need to figure out the area over which the energy gets spread, and thus the area of a sphere.

As a reminder, the invariant distance equation in a homogeneous and isotropic Universe can be written as:

$$ds^2 = -c^2dt^2 + a^2(t)\left[\frac{dr^2}{1-kr^2} + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right)\right] \tag{1.6.1}$$

Box 1.6.1

Calculate the area of a sphere ignoring effects of expansion, in 5 steps.

Exercise 6.1.1: According to the invariant distance equation, what is the distance between (t, r, θ, ϕ) and $(t, r, \theta + d\theta, \phi)$?

Answer

Constant t, r, and ϕ so we have

$$ds = a(t)r\,d\theta$$

Exercise 6.1.2: What is the distance between (t, r, θ, ϕ) and $(t, r, \theta, \phi + d\phi)$?

Answer

Similar to 7.1.1 above, we now have constant t, r, and θ, which gives

$$ds = a(t)r\sin\theta\,d\phi$$

Exercise 6.1.3: What is the area of a rectangle formed with those two lengths?

Answer

This is simply $a^2(t)r^2\sin\theta\,d\theta d\phi$.

Exercise 6.1.4: What is the area of a sphere at coordinate value r with center at the origin?

Answer

Using the result from 7.1.3 above, we now just integrate over θ and ϕ, so the area is

$$A = \int_0^{2\pi}\int_0^{\pi} a(t)rsin\theta d\theta d\phi$$

$$= \int_0^{2\pi} 2a^2(t)r^2 d\phi$$

$$= 4\pi a^2(t)r^2$$

Exercise 6.1.5: Neglecting effects due to expansion (the changing of $a(t)$), how are luminosity and flux related for an observer at the origin and an object at coordinate distance $r = d$? You should find that $F = L/(4\pi d^2 a^2)$.

Answer

We know that Luminosity = (Flux)×(Surface Area). Making the appropriate substitutions we do indeed find that $F = L/(4\pi d^2 a^2)$.

Box 1.6.2

Now let us include effects of expansion. There are two distinct effects here:

Exercise 6.2.1: Convince yourself that the rate of photon arrival is slower than the rate of photon departure by a factor of $1+z$. (Hint: recall the arguments we made in chapter 6 about how the time in between emission of pulses is related to the time in between reception of pulses.)

Answer

As we found in Chapter 6, the rate of arrival of the wave crests will be slower than the rate of emission by the factor of $a(t_r)/a(t_e)$. The same argument applies to the rate of arrival of photons. We also saw that wavelength would be stretched out by a factor $1+z \equiv \frac{\lambda_r}{\lambda_e} = a(t_r)/a(t_e)$. Therefore the rate of arrival of photons will be slowed down by a factor of $1+z$.

Exercise 6.2.2: Convince yourself that the energy of each photon decreases by a factor of $1+z$.

Answer

The relationship between photon energy and wavelength is $E = \frac{hc}{\lambda}$. Substituting this into the definition of redshift, we find that

$$E_r/E_e = \frac{1}{1+z}$$

Each of these two effects reduces the flux by a factor of $1+z$ so the effect of expansion is to alter the flux-luminosity-distance relationship so that:

$$F = \frac{L}{4\pi d^2 a^2 (1+z)^2} \tag{1.6.2}$$

The presence of a here in this result raises a question, which we address next.

Now that the universe is expanding, what value of a should we include in Equation 1.6.2? To answer this, recall that $4\pi d^2 a^2$ is the area of the sphere over which the luminosity is spread, so we can determine power *per unit area*. We should therefore use the value of a at the time the measurement is being made. If we assume the measurement is being made in the current epoch (that we often just simply call "today") and we choose the common convention of normalizing the scale factor so that it's value is one today then we get:

$$F = \frac{L}{4\pi d^2 (1+z)^2} \tag{1.6.3}$$

In Minkowski spacetime (a non-expanding homogeneous and isotropic spacetime) we have the relationship $F = L/4\pi d^2$ where d is the distance between the observer and the source. This relationship motivates the definition of what is called "luminosity distance", d_{lum}, defined implicitly via:

$$F = \frac{L}{4\pi d_{\text{lum}}^2} \tag{1.6.4}$$

Box 1.6.3

Exercise 6.3.1: Make an explicit definition of d_{lum} by solving Equation 1.6.4 for d_{lum}.

Answer

Solving Equation 7.4 for d_{lum} we get

$$d_{\text{lum}} = \sqrt{\frac{L}{4\pi F}}$$

Exercise 6.3.2: In an FRW spacetime, for an observer at the origin, what is the luminosity distance to an object at coordinate distance $r = d$ with redshift z?

Answer

Substituting in Equation 7.3 to our result above we get

$$d_{\text{lum}}^2 = \frac{L}{4\pi}\frac{4\pi d^2(1+z)^2}{L} \quad \Longrightarrow \quad d_{\text{lum}} = d(1+z)$$

So now we know how to infer a particular kind of distance from an observation of a standard "candle." If a source of light is considered a standard candle, that means we know its luminosity. We can measure flux because it's a local property (how much energy per unit time per unit area is flowing past us right here and now). With L and F known we can calculate the luminosity distance. In the next chapter we work out the relationship of luminosity distance and redshift with the history of the scale factor, $a(t)$.

Summary

1. A common convention, that we will adopt, is to normalize the scale factor so that it is equal to one today. If we refer to the current epoch as $t = t_0$ then our normalization choice can be written as $a(t_0) = 1$.
2. The flux, F, from an isotropic emitter with luminosity, L, a coordinate distance d away from an observer observing today, is given by

$$F = \frac{L}{4\pi d^2(1+z)^2}$$

   where z is the redshift of the light from the source due to the expansion.
3. The luminosity distance, d_{lum} from here and now to a source is, by definition, the distance that gives the Euclidean, non-expanding result:

$$F = \frac{L}{4\pi d_{\text{lum}}^2}.$$

   Given summary item (2) above, we see that $d_{\text{lum}} = d \times (1+z)$.

This page titled 1.6: Distances as Determined by Standard Candles is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

- **Current page** by Lloyd Knox is licensed CC BY 4.0.
- **2.7: S07. Distances as Determined by Standard Candles - SOLUTIONS** by Lloyd Knox is licensed CC BY 4.0.

1.7: The Distance-Redshift Relation

For a given scale factor history, $a(t)$, one can work out a relationship between luminosity distance and redshift. This will be useful to us because it indicates how we can *infer* $a(t)$ from measurements of luminosity distance and redshift, over a range of redshifts.

Recall that for light world lines (paths through spacetime), $ds^2 = 0$. For a radial trajectory (one with $d\phi = d\theta = 0$) we thus have $c^2dt^2 = a^2(t)dr^2/(1-kr^2)$. Taking the square root, and choosing the sign so that the photon is headed toward the origin ($dr/dt < 0$) we have:

$$cdt = -\frac{a(t)dr}{\sqrt{1-kr^2}} \tag{1.7.1}$$

Assuming $a(t)$ we could do the integrals on both sides and find out how long it takes for the light to go from $r = d$ to the observer at the origin. But that time interval is not something we can measure, so we'd have a prediction following from the assumed $a(t)$ but no way to confirm it (at least not from what we've developed so far in our exposition here.) What we *can* measure is redshift, which as we've seen depends on the scale factor at the time of emission, so instead we swap out the dt for da and integrate over da. Since $dt = da/(da/dt)$ we get:

$$\frac{cda}{a\dot{a}} = -\frac{dr}{\sqrt{1-kr^2}} \tag{1.7.2}$$

or

$$c\int_{a_e}^{1}\frac{da}{a\dot{a}} = -\int_{d}^{0}\frac{dr}{\sqrt{1-kr^2}} = \int_{0}^{d}\frac{dr}{\sqrt{1-kr^2}} \tag{1.7.3}$$

It is conventional, and we will later find it convenient, to define the Hubble parameter $H \equiv \dot{a}/a$. This is a generalization of the Hubble constant, $H_0 = H(t_0)$ where t_0 is the time today. With this definition we can write:

$$c\int_{a_e}^{1}\frac{da}{a^2H} = \int_{0}^{d}\frac{dr}{\sqrt{1-kr^2}} \tag{1.7.4}$$

Let's work out the consequences of the above in a simple case valid for short travel times and a Euclidean geometry. Putting it more precisely, let's assume $k = 0$ and take $a(t)$ as given by its first order Taylor expansion about the current epoch so that:

$$a(t) = 1 + (t-t_0)\dot{a}|_{t_0} \tag{1.7.5}$$

where for the last term we've indicated it's to be evaluated at time $t = t_0$ (consistent with our assumption of a Taylor expansion). Note that truncating this Taylor expansion to first order means that $da/dt = \dot{a}|_{t_0}$ is a constant. Since the scale factor is unity today (by convention) we also have $\dot{a} = H_0$ and $H \equiv \dot{a}/a = H_0/a$.

Box 1.7.1

Exercise 7.1.1: Plugging $H = H_0/a$ into Equation 1.7.4 one can now do the integral on the left-hand side. The right-hand side could not be easier (since we are assuming $k = 0$). Check that you find:

$$\frac{c}{H_0}\left[\ln(1) - \ln(a_e)\right] = d$$

Answer

$$c\int_{a_e}^{1}\frac{da}{a^2H} = \int_{0}^{d}dr = d$$

$H = \frac{\dot{a}}{a}$ and $\dot{a} =$ constant, so we have $H = \frac{H_0}{a}$. Now,

$$d = \frac{c}{H_0}\int_{a_e}^{1}\frac{da}{a} = \frac{c}{H_0}\Big[\ln(1) - \ln(a_e)\Big]$$

Box 1.7.2

Exercise 7.2.1: Relate a_e to the redshift z, and take advantage of $\ln(1+x) = x$ to first order in x to derive $cz = H_0 d$ to first order in z. How is d here related to luminosity distance? Simplify your result, again assuming $z << 1$. You should find $cz = H_0 d_{\text{lum}}$. Finally, if z is replaced with $z = v/c$ we get Hubble's Law.

Answer

$$\frac{c}{H_0}\Big[-\ln(a_e)\Big] = d$$

$$d = \frac{c}{H_0}\ln\left(\frac{1}{a_e}\right)$$

$$d = \frac{c}{H_0}\ln(1+z)$$

$$d \approx \frac{c}{H_0}z$$

Now multiply both sides by $(1+z)$,

$$d \times (1+z) = \frac{c}{H_0}z(1+z)$$

$$d_{\text{lum}} = \frac{c}{H_0}z(1+z)$$

Since $z << 1$ we can simplify our result to $cz = H_0 d_{\text{lum}}$.

Summary

1. The Hubble parameter is $H \equiv \frac{\dot{a}}{a}$. What we call "the Hubble constant", H_0 is the Hubble parameter evaluated today, $H_0 = H(t_0)$.
2. Luminosity distance and redshift are two things we can measure. The relationship depends on $a(t)$ and the curvature k. In principle, if we measure distances and redshifts for objects at a variety of distances we could then infer $a(t)$ and k. The general relationship between redshift and luminosity distance is contained in these equations:

$$c\int_{a_e}^{1}\frac{da}{a^2 H} = \int_0^d \frac{dr}{\sqrt{1-kr^2}} \tag{1.7.6}$$

and

$$d_{\text{lum}} = d(1+z) \tag{1.7.7}$$

with $1+z = 1/a_e$.

3) For small redshifts, the above reduces to $cz = H_0 d$ for $k = 0$, (and for non-zero k: $cz = H_0\int_0^d \frac{dr}{\sqrt{1-kr^2}}$). If one sets $v = cz$ (which makes sense for a Newtonian interpretation of the redshift), then we arrive at Hubble's Law $v = H_0 d$.

HOMEWORK Problems

Problem 1.7.1

Assume the Hubble parameter varies with scale factor as $H = H_0 a^{-3/2}$ and that $k = 0$. As we will see in subsequent chapters this is what one gets (when $k = 0$) for a universe filled with non-relativistic matter and nothing else. Note that we are using our convention that the scale factor today is unity; i.e., $a(t_0) = 1$ (and further note that we will not continue to give this reminder). Show that the luminosity distance is related to redshift via:

$$d_{\text{lum}} = \frac{2c}{H_0}\left[1-\sqrt{\frac{1}{1+z}}\right] \times (1+z)$$

Problem 1.7.2

Show that to first order in z the above relationship reduces to $cz = H_0 d_{\text{lum}}$; i.e., Hubble's Law.

Problem 1.7.3

Assume the Hubble parameter varies with scale factor as $H = H_0 a^{-1}$ and that $k < 0$. As we will see in subsequent chapters this is what one gets for a universe filled with nothing. Show that

$$d_{\text{lum}} = \frac{1+z}{\sqrt{|k|}} \sinh\left[\sqrt{|k|}\frac{c}{H_0}\ln(1+z)\right].$$

Problem 1.7.4

Use appropriate Taylor expansions to show, once again, that to first order in z the result in 1.7.3 reduces to $cz = H_0 d_{\text{lum}}$.

Problem 1.7.5

Make a qualitative sketch, on the same graph, of d_{lum} vs. z for the universe model in problem 1.7.1 and for the universe model in problem 1.7.3. Assume the same value of H_0 for each. At low z the two curves should be coincident. I just want to see, from your drawings, which one starts to have d_{lum} grow more rapidly with z once z gets big enough that the Taylor series approximations break down. It would be sufficient to look at behavior as $z \to \infty$. To do so, you will want to use $\sinh(x) = (e^x - e^{-x})/2 \to e^x/2$ for large x. Be sure to label your curves.

This page titled 1.7: The Distance-Redshift Relation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

- **Current page** by Lloyd Knox is licensed CC BY 4.0.
- **2.8: S08. The Distance-Redshift Relation - SOLUTIONS** by Lloyd Knox is licensed CC BY 4.0.

1.8: Dynamics of the Expansion

Introduction to the Dynamics chapters, 10-16.

In the following set of chapters we will derive the dynamical equations that relate the matter content in a homogeneous and isotropic universe to the evolution of the scale factor over time. There are fundamentally two independent dynamical equations. The first of these is the Friedmann equation:

$$H^2 = \frac{8\pi G\rho}{3} - \frac{k}{a^2} \tag{1.8.1}$$

where $H \equiv \dot{a}/a$, G is Newton's gravitational constant, k is the curvature constant we have already seen in the FRW invariant distance rule. For simplicity here we have assumed a single type of matter fills the universe; ρ is its mass density.

Note

The generalization could not be more straightforward. For $i = 1$ to N components, replace ρ with $\Sigma_{i=1}^{N}\rho_i$ where ρ_i is the mass/energy density of the ith component.

The second equation tells us how ρ evolves with time (or scale factor) and can be taken to be:

$$a\frac{d\rho}{da} = -3\left(\rho + P/c^2\right) \tag{1.8.2}$$

where P is the pressure.

One might think we need general relativity to derive these equations, and one would be correct. However, we can get surprisingly close with a Newtonian analysis. Here we will follow a Newtonian analysis, and clarify where we are taking something from GR without proof.

We know that a Newtonian analysis of gravitational dynamics will lead to highly accurate predictions when relative speeds are slow and gravitational fields are weak; Newtonian theory follows from general relativity in these limits. Our Newtonian analysis takes advantage of this fact. We can study the dynamics over as small a distance as we like, over which relative speeds are as small as we want, and gravitational potential differences are arbitrarily small. As we will see, the results we achieve by doing so are *independent* of the size we assume. If our results are describing one small region of the universe accurately, by the assumed homogeneity of the spacetime they must be describing all other small regions of the universe accurately, and hence the whole universe accurately.

Our approach risks confusion because we will be moving back and forth between Newtonian descriptions of spacetime and relativistic descriptions. So consider yourself duly warned. Be vigilant and seek clarification when something seems amiss.

This page titled 1.8: Dynamics of the Expansion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.9: A Newtonian Homogeneous Expanding Universe

Let's consider a homogeneous and isotropic expanding universe in a Newtonian manner: space is fixed, and the universe is filled with a fluid that is moving in such a manner as to keep the density spatially uniform. Is such a fluid flow possible? We are going to begin by assuming the fluid is undergoing a uniform expansion so we have distances between all pairs of fluid elements scaling up over time with some scale factor $a(t)$. We will show that the fluid flow resulting from such an assumption is consistent with both the maintenance of uniform density over time and the conservation of mass. So such a flow is indeed kinematically possible. In the next section we will study the dynamics of this flow under the influence of gravity.

With uniform expansion we can write the location of all fluid elements as evolving over time as

$$\vec{x}(t) = a(t)\vec{x}_c \tag{1.9.1}$$

where $\vec{x}_c$ is some fixed quantity for each fluid element. The 'c' stands for comoving. We consider any fluid element following this rule to be 'comoving' with the expansion; i.e., its only motion is due to the expansion. In our completely homogeneous universe, all fluid elements are 'comoving' in this way.

The velocity of any fluid element is then

$$\vec{v} = d\vec{x}/dt = \dot{a}\vec{x}_c = (\dot{a}/a)\vec{x} \tag{1.9.2}$$

and, as expected for uniform expansion, we see we have Hubble's law (or rather a vector version of it).

Does this pattern of fluid flow preserve homogeneity? Let's see by looking at the continuity equation, which follows from mass conservation:

$$\partial\rho/\partial t = -\vec{\nabla}\cdot(\rho\vec{v}). \tag{1.9.3}$$

If we assume ρ is initially uniform then the right-hand side can be written as - $\rho\vec{\nabla}\cdot\vec{v}$. Substituting in the expression for the velocity field in Equation 1.9.2 one can show (see the Exercise) that $\vec{\nabla}\cdot\vec{v} = 3\dot{a}/a$. By assumption, $\dot{a}/a$ is independent of location so the continuity equation says that if the field is initially homogeneous, its time derivative is independent of location so it will stay homogeneous. So we have demonstrated what we wanted to demonstrate: the uniform expansion represented by Equation 1.9.1 leads to a fluid flow that indeed preserves homogeneity.

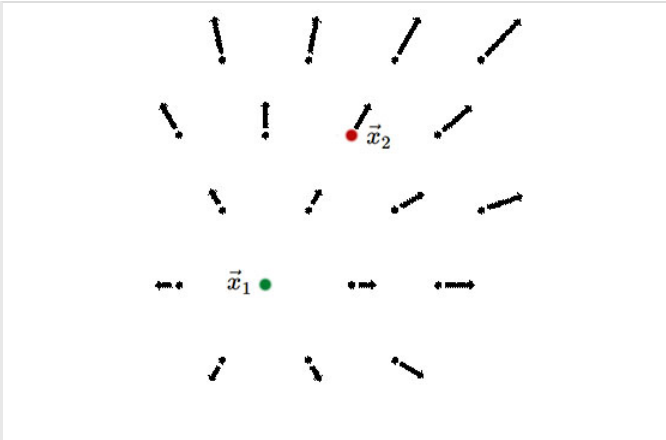
Box 1.9.1

Exercise 10.1.1: Show that if $\vec{x}(t) = a(t)\vec{x}_c$ that $\vec{v} = d\vec{x}(t)/dt \propto \vec{x}$. Find the proportionality constant in terms of $a(t)$ and its time derivative (where by constant here we mean the same everywhere, not necessarily constant in time). Then show that $\vec{\nabla}\cdot\vec{v} = 3\dot{a}/a$.

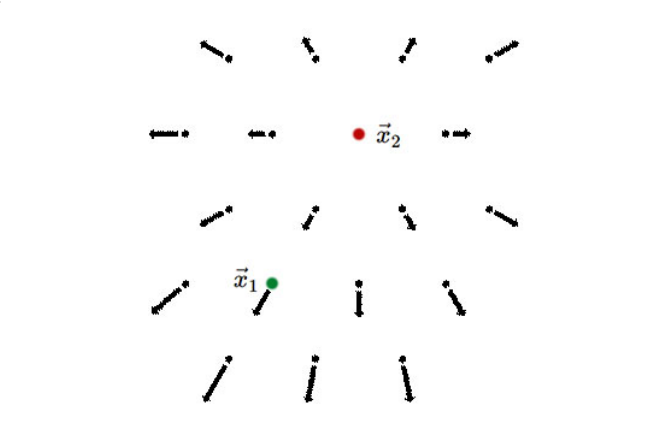
Answer

Since $\vec{x}(t) = a(t)\vec{x}_c$ with $\vec{x}_c$ a constant, we have $\vec{v} = dx(t)/dt = \dot{a}\vec{x}_c = (\dot{a}/a)\vec{x}$ which is Hubble's law with the proportionality between velocity and distance given by $\dot{a}/a$. For the second part, note that $\vec{\nabla} = \hat{x}\partial/\partial x + \hat{y}\partial/\partial y + \hat{z}\partial/\partial z$ and $\vec{v} = (\dot{a}/a)(x\hat{x} + y\hat{y} + z\hat{z})$ so $\vec{\nabla}\cdot\vec{v} = (\dot{a}/a)\times(1+1+1) = 3(\dot{a}/a)$. (Here we've used e.g. $\hat{x}$ as the unit vector in the x direction.)

Hubble's law, as seen in the above exercise, does not look homogeneous. After all, there's a special point that has zero velocity. But this is just a matter of the choice of coordinate system and is not a real inhomogeneity. We'll explicitly demonstrate now that Hubble's law is valid even if you chose a different point to be the one at rest. If the point at rest is $\vec{x}_1$, then any point $\vec{x}$, has velocity $\vec{v} = (\vec{x} - \vec{x}_1)\,\dot{a}/a$. We'll call this the vector version of Hubble's Law. An observer at rest with respect to the fluid at $\vec{x}_1$ will see a flow as shown in the left panel of the figure below.



Coordinate system with $\vec{x}_1$ as the point at rest and seeming center of expansion. Image by Bryan Miller.



Coordinate system with $\vec{x}_2$ as the point at rest. Image by Bryan Miller.

What will an observer at rest with respect to the fluid at location x_2 see? In our original unprimed coordinate system we have

$$\vec{v}_2 = (\vec{x}_2 - \vec{x}_1)\dot{a}/a \quad \text{(unprimed system)} \tag{1.9.4}$$

and we want $\vec{v}_2' = 0$ via a Galilean transformation to the unprimed system, so we subtract $\vec{v}_2$ from the unprimed velocities everywhere:

$$\begin{aligned} \vec{v}' &= \vec{v} - \vec{v}_2 \quad \text{(Galilean velocity transformation rule)} \\ &= (\vec{x} - \vec{x}_1)\dot{a}/a - (\vec{x}_2 - \vec{x}_1)\dot{a}/a \\ &= (\vec{x} - \vec{x}_2)\dot{a}/a \end{aligned} \tag{1.9.5}$$

So we see that we still have Hubble's Law. We see that arranging the fluid so that it all flows away from a given point with speed linearly dependent on distance preserves homogeneity and isotropy.

Although we first derived Hubble's Law in the context of a relativistic description of spacetime, we now see it arising in a Newtonian context. Fundamentally, Hubble's Law follows from the uniformity of the expansion, whether that's a fluid that's expanding or space itself.

HOMEWORK Problems

Problem 1.9.1

Demonstrate that a universe obeying a *nonlinear* version of Hubble's law would *violate* homogeneity. Take $v = H_0 d^2$ for specificity and work in one dimension for simplicity. Show that assuming this velocity pattern about one point leads to a different velocity pattern around other points, thereby demonstrating the violation of homogeneity. You can do this by thinking of 3 points all in one line, with the central point the same distance from the surrounding two.

This page titled 1.9: A Newtonian Homogeneous Expanding Universe is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

- **Current page** by Lloyd Knox is licensed CC BY 4.0.
- **2.9: S10. A Newtonian Homogeneous Expanding Universe - SOLUTIONS** by Lloyd Knox is licensed CC BY 4.0.

1.10: The Friedmann Equation

Sticking with our Newtonian expanding universe, we will now derive the *Friedmann equation* that relates the mass/energy density to the rate of change of the scale factor.

We will proceed by using the Newtonian concept of energy conservation. (You may be surprised to hear me call this a Newtonian concept, but the fact is that energy conservation does not fully survive the transition from Newton to Einstein).

Box 1.10.1

Assume the universe is filled with a fluid with mass density ρ, that is flowing in a manner consistent with Hubble's law. Consider a test particle of mass m a distance $a(t)\ell$ away from the origin of the coordinate system, that moves along with the fluid; i.e., all of its motion relative to the origin is due to the changing of the scale factor. We take the origin of the coordinate system to be at rest.

Exercise 10.1.1: Express the test particle's kinetic energy as a function of $\dot{a}, \ell$ and m.

Answer

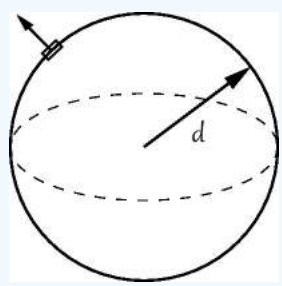
K.E. is $\frac{1}{2}mv^2$, where $v = \dot{a}\ell$, so the test particle's kinetic energy is

$$\frac{1}{2}m\dot{a}^2\ell^2$$

Exercise 10.1.2: Calculate the test particle's potential energy. Do so by considering just the mass interior to a sphere centered on the origin, with radius $a(t)\ell$, as is justified for spherical mass distributions in Newtonian mechanics.

Answer

P.E. is $-\frac{GM(<d)m}{d}$, where by $M(<d)$ we mean the mass contained in the sphere of radius d, so $M(<d) = \frac{4}{3}\pi d^3\rho$.



Therefore the test particle's potential energy is

$$-G\frac{4}{3}\pi\rho d^2 m$$

Exercise 10.1.3: Add these two together and set them equal to a constant. Call the constant κ.

Answer

$$\frac{1}{2}m\dot{a}^2\ell^2 - G\frac{4}{3}\pi\rho d^2 m = \kappa$$

Exercise 10.1.4: Manipulate your equation from Exercise 11.1.3 to get:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\rho}{3} + \frac{2\kappa}{ml^2} \times \frac{1}{a^2}$$

Answer

Recall that $d = a\ell \implies \ell = \frac{d}{a}$, substituting this in and rearranging our equation we get

$$\frac{1}{2}md^2\frac{\dot{a}^2}{a^2} - \frac{1}{2}md^2\frac{8\pi G\rho}{3} = \kappa$$

dividing through by $\frac{1}{2}md^2$ gives

$$\frac{\dot{a}^2}{a^2} - \frac{8\pi G\rho}{3} = \frac{2\kappa}{m}\frac{1}{d^2}$$

Then we substitute back in $d = a\ell$ and solve for $\left(\frac{\dot{a}}{a}\right)^2$:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\rho}{3} + \frac{2\kappa}{ml^2} \times \frac{1}{a^2}$$

Note that everything in that last term in front of the $\frac{1}{a^2}$ factor is a constant in time. Because the other two terms in the equation are constant in space, we can conclude that the third term is also constant in space. Since $\frac{1}{a^2}$ is also constant in space, the combination $2\kappa/(ml^2)$ that precedes it must also be constant in space. We are free to simplify this term by introducing a new space-time constant that we will call $-k$ so the Friedmann equation becomes:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\rho}{3} - \frac{k}{a^2}. \tag{1.10.1}$$

In general relativity this constant, k, is the same as the curvature constant in the FRW invariant distance equation. Of course we cannot use Newtonian physics to derive this fact, as curvature is a non-Newtonian concept. We simply state it here with no proof.

To fully have a handle on the dynamics, what's left is to determine how ρ changes as a changes, a subject we will address in the next few chapters. For now, to explore some consequences of the Friedmann equation, we assume $\rho \propto a^{-3}$. This is a model that most cosmologists, over 25 years ago, thought was probably an excellent approximation to reality. With this assumption, there becomes a connection between the geometry of space, and the fate of the universe. If $k > 0$ the expansion eventually halts and becomes contraction. If $k < 0$ or if $k = 0$ the expansion rate remains positive, approaching but never reaching zero.

Note that the Friedmann equation gives us another means of inferring the curvature constant k. All we need to do is determine the expansion rate today, H_0 and the density of matter today, ρ_0 and plug them in to the Friedmann equation and solve for k. We've already seen, in principle, how to infer H_0 by measuring luminosity distances and redshifts for objects at low redshift and using $cz = H_0 d_{\text{lum}}$. Measuring ρ_0, is trickier, and a subject we may not get to in this course. However, cosmologists have been attempting to measure ρ_0 for decades, in order to then infer k. Because they also used to believe $\rho \propto a^{-3}$, this meant a density measurement could be used to infer the fate of the universe. This situation is the origin of the statement "density is destiny." As we will see, the surprising discovery of cosmic acceleration ($\ddot{a} > 0$) destroyed this notion.

HOMEWORK Problems

Note: Unless directed otherwise, in these first 3 problems assume that $k = 0$ and $\rho \propto a^{-3}$.

Problem 1.10.1

a. What is the age of the universe as a function of H?
b. If there were some non-zero curvature, would the magnitude of its contribution to the expansion rate, relative to that of the energy density, increase or decrease over time?

Problem 1.10.2

What is the furthest physical distance a signal could travel if it started at the beginning? Express your answer as a function of H. To be clear, the distance I mean is the physical distance, at the time when the expansion rate is H, between the particle's actual spatial location and the spatial location it would have had if it had been stationary since the beginning.

Problem 1.10.3

Is there any limit to the comoving distance we can send a signal, if we send one now and are willing to wait an arbitrarily long amount of time? This is equivalent to asking, are there galaxies out at sufficient distance, that a light signal we send to them now will never get to them, even given an infinite amount of time?

Problem 1.10.4

Show that if $\rho \propto a^p$ with $p < -2$, then $\ddot{a} < 0$. Do not assume $k = 0$.

This page titled 1.10: The Friedmann Equation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

- **Current page** by Lloyd Knox is licensed CC BY 4.0.
- **2.10: S11. The Friedmann Equation - SOLUTIONS** by Lloyd Knox is licensed CC BY 4.0.

1.11: Particle Kinematics in an Expanding Universe - Newtonian Analysis

Imagine laying down a grid that gives a spatial coordinate everywhere in space, a grid that expands uniformly, as the uniform fluid filling space also expands uniformly. Let's call those coordinates $\vec{x}$. As the expansion occurs, we consider the origin of the grid ($\vec{x} = 0$) to be at rest. The physical distance from the origin to a point labeled by $\vec{x}$ is $\vec{y} = a\vec{x}$. We call $\vec{x}$ a *comoving* coordinate system. Everywhere the fluid of uniform density has $\dot{\vec{x}} = 0$ and $\dot{\vec{y}} = \dot{a}\vec{x}$. Everywhere, there is a local rest frame, defined by the rest frame of the local part of the grid. Of course this rest frame is different depending on where you are -- after all, every part of the grid is moving away from every other part of the grid.

Now let's consider a particle that is moving with respect to the comoving grid, so that its velocity with respect to the origin has two different contributions. Following from $\vec{y} = a\vec{x}$ and the chain rule we have:

$$\dot{\vec{y}} = \dot{a}\vec{x} + a\dot{\vec{x}} \tag{1.11.1}$$

where the first term on the left-hand side is what we get from the expansion alone, and the second term arises from the particle's motion with respect to the local rest frame. We call this latter term the *peculiar* velocity, $\vec{v}_{\text{pec}}$.

We want to understand how an observer, at rest in their local rest frame, will observe the evolution of peculiar velocities of free particles. In a Newtonian analysis, the local rest frame will be an inertial frame (one in which Newton's laws of motion apply) only if $\dot{a}$ is a constant. With $\dot{a}$ constant, as long as $\vec{y}$ is an inertial frame, then the local frame at $\vec{x}$ will be inertial as well since they only differ by a constant velocity $\dot{a}\vec{x}$. So we will perform our analysis assuming $\ddot{a} = 0$. The more general result is an interesting one that we will comment upon later.

Let's first think visually and qualitatively about how peculiar velocities of free particles will evolve. Imagine a particle that is moving away from the origin faster than its Hubble velocity. It will soon have moved out to larger comoving radial distance, thereby increasing the Hubble velocity. Since it's total velocity is constant (it is a free particle), its peculiar velocity must have decreased.

Box 1.11.1

Exercise 12.1.1: Argue similarly what happens to the magnitude of the peculiar velocity if it starts out negative; i.e., if the particle is decreasing its comoving separation from the origin.

Answer

If the peculiar velocity is negative then the particle, over time, ends up with a smaller distance from the origin then it would have had if its peculiar velocity had been zero. In other words, it starts falling behind. Falling behind naturally brings its motion to be more similar to the surrounding fluid, as the fluid closer to the origin is moving more slowly. This improved agreement between its velocity and that of the surrounding fluid flow can be interpreted as a reduction in the magnitude of the peculiar velocity.

From this qualitative analysis, we expect to find that peculiar velocities decrease over time as a result of the expansion. Let's now do the calculation. Starting from Equation 1.11.1, and recalling our assumption that $\ddot{a} = 0$ we get:

$$\ddot{\vec{y}} = a\ddot{\vec{x}} + \dot{a}\dot{\vec{x}} + \dot{a}\dot{\vec{x}} \tag{1.11.2}$$

From $\vec{v}_{\text{pec}} = a\dot{x}$ one can easily show the above can be rewritten as:

$$\ddot{\vec{y}} = \dot{\vec{v}}_{\text{pec}} + (\dot{a}/a)\vec{v}_{\text{pec}} \tag{1.11.3}$$

Since this is a free particle we have $\ddot{\vec{y}} = 0$ and therefore find:

$$\dot{\vec{v}}_{\text{pec}} = -(\dot{a}/a)\vec{v}_{\text{pec}} \tag{1.11.4}$$

which has the solution $\vec{v}_{\text{pec}} \propto 1/a$. We do indeed find that peculiar velocities of free particles decrease as the universe expands.

Box 1.11.2

Exercise 12.2.1: Fill in the steps in the above derivation.

Answer

Let's look at just one of the three spatial components of the velocity. Notationally, we'll drop the subscript "pec" to make room for the subscript "x" indicating we are talking about the x-component of the velocity. For the other components it is the same solution.

From $dv_x/dt = -(da/dt)/av_x$ we can rearrange and cancel terms to get $dv_x/v_x = -da/a$, which we can integrate up:

$$\int_{v_{x,i}}^{v_x} \frac{dv'_x}{v'_x} = -\int_{a_i}^{a} \frac{da'}{a'}$$

where the i subscript indicates "initial" and the integrals are easily done to get

$$\ln(v_x/v_{x,i}) = -\ln(a/a_i) = \ln(a_i/a)$$

which can be solved to find

$$v_x = v_{x,i}(a_i/a).$$

The same goes for the other components of the peculiar velocity. So we have what we wanted to show, that the peculiar velocity reduces with expansion as $1/a$.

While to get this result ($\vec{v}_{\text{pec}} \propto 1/a$) from our Newtonian analysis we had to make the assumption $\ddot{a} = 0$, in general relativity the result holds even without the assumption. Let's look at this more closely to see what it means.

If we did the above derivation *without* the $\ddot{a} = 0$ assumption we would find instead:

$$\ddot{\vec{y}} = \dot{\vec{v}}_{\text{pec}} + (\dot{a}/a)v_{\text{pec}} + \ddot{a}\vec{x} \tag{1.11.5}$$

This additional term makes perfect sense in the Newtonian theory. If the scale factor is accelerating, then a particle at a fixed value of $\vec{x}$ will be accelerating at a rate $\ddot{a}\vec{x}$. Due to this acceleration, the local rest frame around $\vec{x}$ is an accelerating frame, not an inertial one. If no force is applied (so $\ddot{y} = 0$), a particle at $\vec{x}$ will obey $\dot{\vec{v}}_{\text{pec}} = -\ddot{a}\vec{x} - (\dot{a}/a)v_{\text{pec}}$. From this one can see that if the peculiar velocity is initially zero, it will become negative (headed toward the origin), as the local grid accelerates away from it.

While in the Newtonian theory a free particle has $\ddot{\vec{y}} = 0$, in the Einstein theory the local fluid rest frame *is* a locally inertial frame even in the presence of $\ddot{a} \neq 0$. So a free particle, rather than obeying $\ddot{\vec{y}} = 0$ everywhere, will obey $\ddot{\vec{y}} - \ddot{a}x$ = 0 everywhere. Thus even with the addition of this extra term we still find $\dot{\vec{v}}_{\text{pec}} + (\dot{a}/a)v_{\text{pec}} = 0$ and therefore $\vec{v}_{\text{pec}} \propto 1/a$ still holds.

You can actually evaluate for yourselves, using tools we've already presented, the above claim that a particle with fixed value of x is not accelerating even in the presence of $\ddot{a} \neq 0$. By not accelerating, we mean that an accelerometer on this same path through spacetime (constant x) will register zero. Another way of putting it: no force would be required to keep the particle on the path. All you need is to know that in general relativity objects in free fall (those experiencing no acceleration) from event A to event B follow the path that maximizes the proper time. Now you can check that the path from (x_1, t_1) (that we'll call event A) to (x_1, t_2) (that we'll call event B) with x always equal to x_1 maximizes $\int_A^B \sqrt{-ds^2/c^2}$ even if $\ddot{a} = 0$. To do so, you can use the calculus of variations result we described in chapter 1.

In the above we assumed non-relativistic speeds. We will also want to know how particles traveling at speeds near the speed of light, or even at the speed of light, are affected by expansion. You have actually already shown how particles traveling at the speed of light are affected by the expansion, because you have worked out how light redshifts. You'll recall that:

$$\lambda_{\text{received}}/\lambda_{\text{emitted}} = a_{\text{received}}/a_{\text{emitted}} \tag{1.11.6}$$

i.e., $\lambda \propto a$. From the fact that the momentum of a photon p is inversely proportional to its wavelength we have:

$$p \propto 1/a \tag{1.11.7}$$

Note that for non-relativistic particles we also have $p \propto 1/a$ (since $v \propto 1/a$ and $p = mv$). It turns out that this is a generally correct result for free particles in an expanding universe, their peculiar momentum decreases as $1/a$.

This page titled 1.11: Particle Kinematics in an Expanding Universe - Newtonian Analysis is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

- **Current page** by Lloyd Knox is licensed CC BY 4.0.
- **2.11: S12. Particle Kinematics in an Expanding Universe- Newtonian Analysis - SOLUTIONS** by Lloyd Knox is licensed CC BY 4.0.

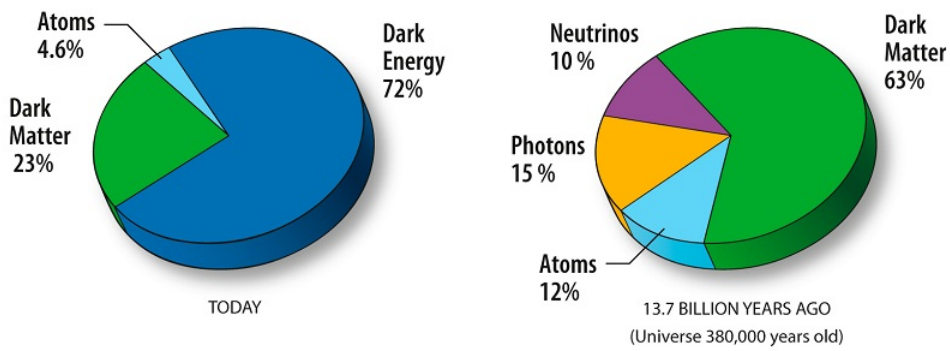
1.12: The Evolution of Mass-Energy Density and a First Glance at the Contents of the Cosmos

We've seen that the rate of change of the scale factor depends on the mass density ρ. In order to determine how the scale factor evolves with time, we thus need to know how the density evolves as the scale factor changes. In this section we'll work that out for three cases. The first two are collections of particles: 1) non-relativistic particles, by which we mean those with rest-mass energy (mc^2) much greater than kinetic energy, and 2) relativistic particles, by which we mean particles with much greater kinetic energy than rest-mass energy. The former we call "matter" and the latter we call "radiation." The third one is much more exotic: a cosmological constant.

These are the three categories we need to describe the evolution of mass-energy density for all the significant components of the standard cosmological model. The relative contributions of these components to the total mass-energy density today, and over 13.7 billion years ago, are shown in the graphic to the right, as cosmologists have estimated them using data from NASA's Wilkinson Microwave Anisotropy Probe (WMAP) satellite. We'll call these components out as we consider Matter, Radiation, and the cosmological constant.

Matter

Let's begin by thinking about how the energy density of the gas in this room would evolve under expansion. Before we even get to that though, let's compare the kinetic energy of a typical Nitrogen molecule in the room to its rest-mass energy. The kinetic energy is roughly given by $k_B T_{\text{room}} \simeq 1/40\text{eV}$ where eV is a unity of energy called an electron Volt equal to the kinetic energy that an electron gains crossing a potential difference of 1 Volt. To get the rest mass, note that Nitrogen's most abundant isotope has 7 protons and 7 neutrons, and the molecule is two Nitrogen atoms so its mass is about 28 times the mass of the proton. The proton mass is almost $1 \times 10^9\text{eV}/c^2$, so the Nitrogen molecule rest mass energy is roughly $mc^2 = 28 \times 10^9\text{eV}$. You can see that its rest mass energy is *much* greater than its kinetic energy.



In thinking about the energy density of a gas of particles we need to keep track of the rest mass energy and the kinetic energy. However, for non-relativistic particles, those moving much more slowly than the speed of light, like the particles in the gas in this room, the kinetic energy is tiny compared to the total energy and we can ignore it. That makes our calculation very straightforward. Expansion will dilute the number of particles by the increase in volume.

Box 1.12.1

Exercise 12.1.1: For a collection of particles all of the same mass, m, their energy density is given by:

$$\rho c^2 = mnc^2$$

where n is the number density of the particles. In your own words, argue that $n \propto a^{-3}$, and hence $\rho \propto a^{-3}$. It might be helpful to track a collection of the particles in a cubic volume over time. Have the volume expand so that it always contains those particles and no others. What happens to the total energy in the box as the scale factor increases? What happens to the volume of the box as the scale factor increases?

Answer

If the particles are not being created or destroyed then the number, N in some comoving volume is fixed. But the volume in the comoving volume is increasing as a^3. The number density $n = N/V$ is thus proportional to a^{-3}.

In the WMAP graphic, "Atoms" and "Dark Matter" are both forms of Matter. Atoms are just the usual matter we are familiar with from everyday life. They are the elements of the periodic table. We actually don't know what the dark matter is. But there are many different observations that can be most easily understood if we assume there is a significant amount of some unknown, not-yet-detected, type of non-relativistic particle that contributes more to the mass-energy of the universe than atoms do by a factor of 5.

Radiation

For any collection of particles the energy density is given by $\rho c^2 = \bar{E}n$ where $\bar{E}$ is the average energy of the particles. For massless particles, $E = pc$ so $E \propto 1/a$ for each particle so we also know that $\bar{E} \propto 1/a$. As long as particles are not being destroyed, just like for the non-relativistic particles, $n \propto 1/a^3$. Putting this together, for relativistic particles we have $\rho \propto a^{-4}$. In the WMAP graphic, Photons and Neutrinos both count as radiation. Photons are the clearest case since, as they have no mass, their kinetic energy is always much greater than their rest-mass energy. Neutrinos are subatomic particles that do have a small amount of mass, but for much of the history of the universe we expect that these particles have much greater kinetic energy than rest-mass energy and hence qualify as radiation.

The Cosmological Constant

It is a logical possibility, consistent with Einstein's equations, that there is an energy density associated with space itself; i.e., a certain amount of energy in every cubic centimeter, an amount that does not change with time. Thus, by definition, for a cosmological constant the mass-energy density is independent of the scale factor: $\rho \propto a^0$. As we will see, there is evidence supporting the existence of a non-zero cosmological constant. Einstein considered this possibility early on as a means to explain why the universe is static (as he thought it was), rather than expanding or contracting, when he introduced the cosmological constant via an additional term in his field equations.

Einstein's reasons for introducing the cosmological constant turned out to be unfounded. In 1929 Edwin Hubble reported his inferences of recession velocity and distance for a set of (relatively nearby) galaxies, that showed a roughly linear trend of increasing velocity with distance, just as one would expect from a uniform expansion. Einstein missed the opportunity to predict the discovery of the expansion of the universe, a missed opportunity he referred to as his greatest blunder.

The cosmological constant though has refused to die. There are two reasons for this. The first is due to the fact that if one tries to calculate, using quantum field theory, the energy density that is in every cubic centimeter of space from the zero-point energy of all the quantum fields, one gets an enormously large energy density, larger than observational limits by about 10^{120}. This huge embarrassment of modern physics is called ``the cosmological constant problem."

The second is that over the past twenty years strong evidence has emerged that the dominant contribution to the mean energy density of the universe in the current epoch is something that is behaving a lot like a cosmological constant. As we will see soon, the observational evidence comes from inferences of the relationship between distance and redshift that indicate the expansion rate is accelerating; i.e., that $\ddot{a} > 0$. Radiation and matter lead to deceleration, while a cosmological constant can produce acceleration (as you will shown in the Box below). The first claims of acceleration from redshift-distance inferences were published in 1998, and were based on observations of Type 1a supernovae. Three of those leading these efforts were awarded the Nobel Prize in Physics in 2011 for their work. The "Dark Energy" label in the WMAP graphic is a more general term than "cosmological constant." It is the general name for the component of the universe that is causing acceleration in the current epoch. A cosmological constant is a very specific kind of dark energy.

Box 1.12.2

Exercise 12.2.1: Show that for an expanding universe with $k = 0$ and only matter or radiation that $\ddot{a} < 0$. Start from the Friedmann equation. (Tip: you do not need to actually calculate $\ddot{a}$; instead, you can show that $\dot{a}$ will decrease as a increases. This may be easier).

Answer

<u>Matter Case:</u>

$$\dot{a}^2 = a^2 \frac{8\pi G\rho_0}{3} a^{-3} = \frac{H_0^2}{a}$$

Since a is increasing with time, H_0^2/a decreases with time, which means $\dot{a}^2$ decreases with time.

$\Longrightarrow$ the magnitude of $\dot{a}$ decreases with time, and since $\dot{a} > 0$ that means $\dot{a}$ decreases with time.

More mathematically:

$$\dot{a} = \frac{H_0}{a^{1/2}} \quad \Longrightarrow \quad \ddot{a} = -\frac{H_0}{2a^{3/2}}\dot{a}$$

Since $\dot{a} > 0$, it implies $\ddot{a} < 0$.

<u>Radiation Case:</u>

$$\dot{a}^2 = a^2 \frac{8\pi G\rho_0}{3} a^{-4}$$

$$\dot{a} = \frac{H_0}{a}$$

$$\ddot{a} = -\frac{H_0}{a^2}\dot{a}$$

Since $\dot{a} > 0$ we see $\ddot{a} < 0$.

Exercise 12.2.2: Show that for an expanding universe with $k = 0$ and only a cosmological constant that $\ddot{a} > 0$. Start from the Friedmann equation.

Answer

$$\frac{\dot{a}^2}{a^2} = \frac{8\pi G\rho_0}{3} a^0 \quad \Longrightarrow \quad \dot{a} = H_0 a$$

Therefore, $\ddot{a} = H_0\dot{a} > 0$.

Summary

The universe has components that change with scale factor in three different ways:

1. Non-relativistic matter, aka "Matter" has a mass-energy density $\rho \propto a^{-3}$,
2. Relativistic matter, aka "Radiation", has a mass-energy density $\rho \propto a^{-4}$, and
3. Dark Energy, whose mass-energy density evolves much more slowly; for the specific case of a cosmological constant $\rho \propto a^0$.

These different behaviors lead to them having different mixes over time in the history of the universe, and explain the differences in the two pie charts in the WMAP graphic. There are still neutrinos and photons around in the current epoch, but their contributions are just so small they have not been included in the graphic. Their steeper dependence on a means that at earlier and earlier times they contributed a greater and greater share of the total mass-energy budget.

HOMEWORK Problems

Note: Unless directed otherwise, in these first three problems assume that $k = 0$ and $\rho \propto a^{-3}$.

Problem 1.12.1

In the standard model of cosmology the universe has gone through at least three different "eras." These are the radiation-dominated era, the matter-dominated era and the dark energy-dominated era. In the radiation-dominated era, for example, more mass-energy comes from radiation than from matter or dark energy. Think about how energy density evolves with scale factor for each of these components and identify which era was first, which second and which third (that is, specify the temporal ordering). Make a sketch of $\log(a)$ vs. $\log(\rho)$ for each of the components, all on the same graph. On the graph indicate the ranges of scale factor for each of the three eras. I am only looking for something qualitatively correct here. Do not worry about having the right value of the scale factor for the transitions, just have the eras in order. Your axes need not have any numbers on them, but do indicate where on the $log(a)$ axis today is.

Problem 1.12.2

From the time referred to in the pie chart to the right in the WMAP graphic, to "TODAY", the universe has expanded by a factor of about 1100. Given that information, and assuming that over this time period there has been negligible destruction or creation of photons, atoms, and dark matter, about what fraction of the mass-energy density today is contributed by photons?

Problem 1.12.3

Later we will study the epoch of "big bang nucleosynthesis" (BBN) when most of the Helium in the universe was created, as well as trace amounts of some other light elements. The scale factor was about one million times smaller in this epoch than it was at the time referred to in the pie chart on the right in the WMAP graphic. Assuming that neutrino mass can be ignored between these two different times, what is the ratio of radiation mass-energy density to matter mass-energy density at the epoch of BBN?

This page titled 1.12: The Evolution of Mass-Energy Density and a First Glance at the Contents of the Cosmos is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

- **Current page** by Lloyd Knox is licensed CC BY 4.0.
- **2.12: S13. The Evolution of Mass-Energy Density and a First Glance at the Contents of the Cosmos - SOLUTIONS** by Lloyd Knox is licensed CC BY 4.0.

1.13: Energy and Momentum Conservation

The lack of energy conservation in an expanding universe is quite surprising to people with any training in physics and therefore merits some discussion, which we present here in this chapter. The student could skip this chapter and proceed to 15 without serious harm. If, subsequently, the lack of energy conservation becomes too troubling, know that this chapter is here for you.

We begin by reminding the reader of the deep connection between symmetries of the action and conserved quantities. For example, if the action (time-integral of the Lagrangian) is invariant under time translations (and hence a symmetry of the action) then energy is conserved. Likewise, if spatial translations do not change the action, then momentum is conserved.

To review, let us consider a single particle moving in one dimension in a possibly space and time-dependent external potential with Lagrangian:

$$L(x, \dot{x}, t) = \frac{1}{2} m \dot{x}^2 - V(x, t) \tag{1.13.1}$$

The action is given by integrating along a trajectory between two points fixed in space and time, points 1 and 2:

$$S = \int_1^2 L(x, \dot{x}, t) dt \tag{1.13.2}$$

Although it is not obvious from the notation, it depends on an assumed $x(t)$. Invariance of the action under time translation means that if we send t to $t + \delta t$, the resulting change in S, δS, will be zero. It should be clear that this will be the case as long as L has no explicit time dependence; i.e., if $\partial L / \partial t = 0$. Likewise, invariance of the action under spatial translation means that if we send x to $x + \delta x$, the resulting change in S, δS, will be zero. This will be the case if $\partial L / \partial x = 0$.

From the Euler-Lagrange equations one can now see directly that space-translation invariance leads to momentum conservation. By definition, the momentum conjugate to x is $p = \partial L / \partial \dot{x}$. For our particular Lagrangian, that gives $p = m\dot{x}$ as expected. The Euler Lagrange equations are:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}}\right) = \frac{\partial L}{\partial x} \tag{1.13.3}$$

which for our Lagrangian becomes:

$$\frac{d}{dt} p = -\frac{\partial V(x, t)}{\partial x} \tag{1.13.4}$$

So we see that p does not change with time if the potential has no dependence on x. Of course! If the potential has no dependence on x then it is not applying any force to the particle.

Seeing the consequences of time-translation invariance takes a little bit more work. However, from the Euler-Lagrange equations one can derive:

$$\frac{d}{dt}\left(\dot{x} \partial L / \partial \dot{x} - L\right) = -\partial L / \partial t \tag{1.13.5}$$

so we see that the quantity after d/dt on the left-hand side is conserved if L has no explicit time dependence. As long as the kinetic energy is quadratic in the velocity, and the potential does not depend on the velocity, then that quantity is equal to the total energy. Thus, given these conditions, energy conservation follows from time translation invariance.

For our Lagrangian the above equation becomes:

$$\frac{d}{dt}\left(\frac{1}{2} m \dot{x}^2 + V(x, t)\right) = \partial V(x, t) / \partial t \tag{1.13.6}$$

Thus we see total energy is conserved if the potential energy function does not have a dependence on t. Note that this is not a statement that the potential energy of the particle cannot change with time (which would be $dV/dt = 0$) but is instead a statement about the functional form of $V(x, t)$, that it cannot have an explicit dependence on time. For example, $V(x) = 1/2kx^2$, has no explicit time dependence ($\partial V / \partial t = 0$) even though x will change with time (so $dV/dt = kx\dot{x} \neq 0$), so for this potential, energy is conserved. However, if the spring constant were to change with time so $V(x) = 1/2k(t)x^2$ then energy would not be conserved. Imagine the behavior of a mass attached to a spring with time-dependent spring constant. Can you see how energy would not be conserved?

Let's now consider how this works in an expanding universe. Consider a free non-relativistic particle of mass m. If we adopt the point of view that the comoving grid everywhere specifies local rest, then we will naturally define kinetic energy based on departures from local rest.This leads us to write the velocity of the particle in terms of $\dot{x}$ so that the velocity is $a\dot{x}$ and the Lagrangian is $L = 1/2ma^2\dot{x}^2$.

One might be tempted to use the physical coordinates $y = ax$ and write $v = \dot{y}$ and $L = 1/2m\dot{y}^2$ and therefore $m\dot{y}$ is conserved. However, this Lagrangian would give kinetic energy to a particle that is stationary with respect to the local rest frame, as long as it is not at the origin. Further, in a universe with \(\ddot a \ne 0\) we run into the same problem as discussed in chapter 12 that the coordinate system is locally inertial near the origin but not globally inertial. The comoving coordinate system, in contrast, is globally inertial.

Let's therefore return to $L = 1/2ma^2\dot{x}^2$ as our free-particle Lagrangian. Note that for this Lagrangian there is no explicit dependence on x so the momentum conjugate to x will be conserved. That momentum is $p = \partial L/\partial\dot{x} = ma^2\dot{x}$ = amv which is a times the usual momentum. Since amv is conserved, we find $mv \propto 1/a$.

Thus, in an expanding homogeneous universe, spatial homogeneity; i.e., translational invariance, leads to a conserved quantity which is the usual momentum divided by the scale factor. Thus we see once again that free particles in an expanding universe have a momentum that decreases as $1/a$.

What about energy, the conserved quantity associated with time translation invariance? We do not have time translation invariance because of the expansion! Consequently, energy is not conserved as one can see explicitly for our free particle from $dE/dt = -\partial L/\partial t = -ma\dot{a}\dot{x}^2 \neq 0$. Of course this result is consistent with our finding that the momentum of the free particle decreases as the universe expands.

I wrote at the beginning of this chapter that the lack of energy conservation in the expanding universe comes as a surprise to many. I ran up against this when I wrote an article for Physics Today about cosmic inflation. In an initial draft I wrote, about the patch that eventually becomes our observable universe, "The total mass-energy in this patch was about 10^4 Joules, the caloric content of two diet Cokes." This line, which I loved, was then removed by the editor! I was very disappointed! The editor explained he was skeptical it was true and was afraid the readers would be as well. I tried to address the skepticism with this suggested replacement:

"The total mass-energy in this patch was about 10^4 Joules, the caloric content of two diet Cokes. (Surprising to most physicists is that there is no global conservation of energy in general relativity, a fact that allows for the existence today of regular Coke, too---as well as the other $\sim 10^{70}$ Joules of mass-energy in the currently observable universe.)"

which had the added benefit of further play on the Coke theme! Fortunately, the editor was open minded and we settled on this for the final published version:

"The total mass-energy in that patch was about 10^4 Joules, the caloric content of two diet Cokes. Today's universe includes regular Cokes and about 10^{70} joules of mass energy from other sources, a result compatible with the perhaps surprising fact that energy is not conserved in general relativity."

I also ran up against it a previous time I taught this class, when students pointed out to me that the textbook we were using for the course explicitly claimed energy was conserved. I pointed it out to the author, who quickly agreed with me it was an error. He was in fact quite unhappy he was finding this out just a little too late to keep the error from propagating into the next edition of his textbook.

This page titled 1.13: Energy and Momentum Conservation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.14: Pressure and Energy Density Evolution

There is a sense in which energy is conserved in general relativity. We say it is *locally* conserved, which effectively means that in a sufficiently small region of spacetime, the change in energy is equal to the flux of energy across the boundary of the region, including that via any work being done on the region. From this principle, or from the Einstein equations themselves, someone with some skill in general relativity can derive for a homogeneous expanding universe that:

$$dE = -P\,dV \tag{1.14.1}$$

where E is the energy in some volume V. This *looks* a lot like conservation of energy as we are used to seeing it. Indeed it is the first law of thermodynamics, in the special case of no heat flow across the boundary. If you have a gas in a volume V and you squeeze it down by dV, the work you do ($-PdV$), increases the kinetic energy (and hence total energy) of the gas particles by

$$dE = -P\,dV. \tag{1.14.2}$$

Since dV is negative, this is an increase in energy, assuming $P > 0$.

But the simplicity of the result is deceptive. As soon as one starts to ask some obvious questions about it, things can become confusing. In a homogeneous expanding universe, how is the work being done? There are no pressure gradients to push things around. Then is it somehow being done by gravitational potential energy? That is a reasonable guess, since the expansion of space is a gravitational effect.

These questions implicitly assume energy is globally conserved, when that is not actually the case in general relativity. We can, however, use our Newtonian intuition to guide us about how the gas will behave given that the region it occupies is expanding. If the volume slowly increases that is containing a gas (with $P > 0$), then the energy of that gas will decrease no matter if it's because of the expansion of space or the expansion of the walls that contain the gas.

From $dE = -PdV$ we can derive how energy density evolves as the scale factor evolves. Gas comoving with the expansion and in a region with comoving volume V_c, occupies a physical volume of a^3V_c. The energy content of this homogeneous gas is $\rho c^2 a^3 V_c$. Thus Equation 1.14.1 leads to

$$a^3c^2d\rho + 3\rho c^2a^2da = -3Pa^2da \tag{1.14.3}$$

or

$$a\frac{d\rho}{da} = -3(P/c^2 + \rho) \tag{1.14.4}$$

Box 1.14.1

Exercise 14.1.1: Use Equation 1.14.4 to find $P(\rho)$ for non-relativistic matter, given that $\rho \propto a^{-3}$.

Exercise 14.1.2: Use Equation 1.14.4 to find $P(\rho)$ for relativistic matter, given that $\rho \propto a^{-4}$.

Exercise 14.1.3: Use Equation 1.14.4 to find $P(\rho)$ for a cosmological constant, given that $\rho \propto a^0$.

Answer

The equation is

$$a\frac{d\rho}{da} = -3(P/c^2 + \rho)$$

Plugging in $\rho \propto a^n$ we get $n\rho = -3(P/c^2 + \rho)$. Solving for P for $n = -3, -4, 0$ we find $P = 0$, $P = \rho c^2/3$, and $P = -\rho c^2$ respectively.

For Exercise 15.1.1 you should find that $P = 0$. This might be surprising since non-relativistic matter in general has non-zero pressure. Remember though that our $\rho \propto a^{-3}$ result came from neglecting all kinetic energy of the gas, because it was so small compared to the kinetic energy. Of course it is not exactly zero so the pressure is also not exactly zero. For 15.1.2 you should find $P(\rho) = \frac{1}{3}\rho c^2$.

Perhaps most surprisingly, for 15.1.3 you should find that the pressure is *negative*. More precisely, you should find that $P(\rho) = -\rho c^2$. How can this be? What does it mean to have negative pressure? We will explore these questions in a homework problem.

We now have a closed set of equations that we can use to solve for the evolution of the scale factor. Given the mass-energy density of all the components of the universe at one time and their equations of state $P(\rho)$, and given the expansion rate at that same time, we can find how these densities evolve and how the scale factor does as well. We summarize the key equations here, as well as writing them out with explicit sums over components for the first time.

$$H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi g \Sigma_i \rho_i}{3} - k/a^2 \tag{1.14.5}$$

and

$$a\frac{d\rho_i}{da} = -3(P_i/c^2 + \rho_i) \tag{1.14.6}$$

where the subscript i enumerates the different contributors to the energy density.

Definitions of some prevalent notation

Densities are often expressed in units of the critical density, with the critical density defined as the total density of a zero-mean-curvature universe with expansion parameter H. Since the Friedmann equation is $H^2 = 8\pi G\rho_{\text{total}}/3 - k/a^2$ this means that the critical density, ρ_{c}, is such that $H^2 = 8\pi G\rho_{\text{c}}/3$. More explicitly

$$\rho_{\text{c}} \equiv 3H^2/(8\pi G). \tag{1.14.7}$$

The symbol used for the density of component i in units of the critical density is $\Omega_i \equiv \rho_{i,0}/\rho_{\text{c},0}$. Often when cosmologists write densities out in terms of Ω they implicitly mean the value in the current epoch. Ideally, we would use a 0 subscript in such cases, in order to explicitly denote the current epoch, but we don't usually do that. So, for example, the Friedmann equation for a universe with pressureless matter, a cosmological constant, radiation, and zero mean curvaturecan be rewritten as

$$H^2(a) = H_0^2\left[\Omega_{\text{m}}a^{-3} + \Omega_\Lambda + \Omega_{\text{rad}}a^{-4}\right]. \tag{1.14.8}$$

We also sometimes use $\Omega_{\text{k}} \equiv -k/H_0^2$.

Box 1.14.2

Exercise 14.2.1: Show that $\Sigma_i\Omega_i + \Omega_k = 1$.

Answer

We have $H^2 = 8\pi G\rho/3 - k/a^2$, $\Omega_i \equiv \rho_{i,0}/\rho_c$, and the critical density today, ρ_c defined indirectly via $H_0^2 = 8\pi G\rho_c/3$. Recall that ρ in the Friedmann equation is the total density so $\rho = \Sigma_i\rho_i$.

Let's take the Friedmann equation, evaluated today (so $H_0^2 = 8\pi G\rho_0/3 - k$ and divide each term by either H_0^2 or $8\pi G\rho_c/3$. We can divide by either because they are equal. We get

$$1 = \Sigma_i\rho_{i,0}/\rho_c - k/H_0^2 = \Sigma_i\Omega_i + \Omega_k \tag{1.14.9}$$

if we also use the given definition of $\Omega_k \equiv -k/H_0^2$.

Another notational convenience sometimes uses is to define h as a way of quantifying the Hubble constant. This "little h" is defined such that

$$H_0 = 100h \text{ km/sec/Mpc}. \tag{1.14.10}$$

Saying "$h = 0.72$" is the same as saying $H_0 = 72$ km/sec/Mpc. I mention this notation because it leads to yet another common way of talking about densities. Sometimes we run across use of, for example, $\Omega_{\text{m}}h^2$, or even ω_{m} which is the same thing by definition. Why would we do this? It's a convenient way of writing out the density, but with no actual dependence on the critical density. In a homework problem you will show how to take a density in units of grams per cubic centimeter and find out what this corresponds to in terms of Ωh^2. You will see it has no dependence on the value of the expansion rate because the h^2 in Ωh^2 cancels the h^2 in the critical density.

Homework

14.1: Imagine you have a box of volume V full of a substance with energy density ρc^2. Outside the box the energy density is zero. Imagine that if you expanded the box by some amount dV that the energy density inside would not drop, but would stay constant.

A) By how much would the energy inside the box increase if this expansion occurred?

B) Imagine you are pulling on the walls of the box to make this increase in volume happen. Would it be hard to do? Would it require work? How much work? Articulate why ascribing $P < 0$ to the material inside the box makes sense.

14.2: For a substance with $P = w\rho c^2$ with w a constant, find n such that $\rho \propto a^{-n}$. Explicitly write out how ρ depends on a for these cases: $w = 0, 1/3, -1, +1$. For example, for $w = 0$ write, "For $w = 0$ we find $\rho \propto a^{-3}$."

14.3: According to multiple lines of argument (one of which we will learn about when we study big bang nucleosynthesis), the mean density of baryonic matter in the universe (matter whose mass comes from protons and neutrons and nuclei made out of protons and neutrons) is such that $\Omega_b h^2$ is about 0.022. What is the mean density of baryons in the universe in grams per cubic centimeter? If $H_0 = 72$ km/sec/Mpc, what is Ω_b?

This page titled 1.14: Pressure and Energy Density Evolution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

- **Current page** by Lloyd Knox is licensed CC BY 4.0.
- **2.13: S15 Pressure and Energy Density Evolution SOLUTIONS** by Lloyd Knox is licensed CC BY 4.0.

1.15: Distance and Magnitude

Distances

We have the invariant distance equation for a homogeneous and isotropic universe (an FRW spacetime):

$$ds^2 = -c^2dt^2 + a^2(t)\left[\frac{dr^2}{1-kr^2} + r^2\left(d\theta^2 + \sin^2(\theta)d\phi^2\right)\right]. \tag{1.15.1}$$

Here we introduce several distance definitions, and how they are related to the coordinate system that leads to the above invariant distance expresson.

Luminosity distance: By definition of luminosity distance d_L,

$$F = \frac{L}{4\pi d_L^2} \tag{1.15.2}$$

which is the relationship we expect in a Euclidean geometry with no expansion, assuming an isotropic emitter. We also calculated the relationship between flux and luminosity in an FRW spacetime and found

$$F = \frac{L}{4\pi r^2(1+z)^2} \tag{1.15.3}$$

so we conclude that in an FRW spacetime, $d_L = r(1+z)$.

Angular diameter distance: By definition of angular-diameter distance, d_A,

$$\ell = \theta d_A \tag{1.15.4}$$

where θ is the angle subtended by an arc of a circle with length ℓ, as it would be measured with measuring tape. By the angle subtended, we mean the angle between two light rays, one coming from one end of the arc, and the other from the other end of the arc. If we place ourselves in the center of the coordinate system we can work out what this means in terms of coordinates. Place the observer at the spatial origin $r = 0$ and at time equals today. Place one end of the arc at $r = d, \theta = 0, \phi = 0$ and the other at $r = d, \theta = \alpha,\ \phi = 0$. Light will travel from both of these points to the origin along purely radial paths; i.e., with no change in θ or ϕ. So the angle they subtend upon arrival is α. We can use the invariant distance expression to work out that $\ell = a\alpha d$ where a is the scale factor at the time the light we are receiving today is emitted from the object. Thus $d_A = \ell/\theta = ad\alpha/\alpha = ad$ where d is the radial coordinate separation between the object and the observer.

Comoving angular diameter distance: This is simply the angular diameter distance divided by the scale factor. We will reserve D_A for comoving angular diameter distance. The comoving angular diameter distance between $r = 0$ and $r = d$ is $D_A = d$.

Box 1.15.1

Exercise 15.1.1: In an FRW spacetime, how are D_A and d_L related?

Box 1.15.2

Exercise 15.2.1: What is the comoving distance, ℓ, from the origin to some point with radial coordinate value r, along a path of constant θ and ϕ?

Curvature integrals: Although we've made use of a first order Taylor expansion to analytically solve the above integral, the exact integral does have an analytic solution. For $k > 0$, $\ell = (1/\sqrt{k})\sin^{-1}(\sqrt{k}r)$. For $k < 0$, $\ell = (1/\sqrt{-k})\sinh^{-1}(\sqrt{-k}r)$.

To work out how the comoving angular diameter distance D_A is related to the scale factor at the time light was emitted, a, we look at how light travels from coordinate value r to the origin. Light has $ds^2 = 0$, and from that we get

$$\int_0^r \frac{dr}{\sqrt{1-kr^2}} = \int_t^{t_0} cdt/a = c\int_a^1 da/(a^2H),$$

$$(1/\sqrt{-k})\sinh^{-1}(\sqrt{-k}r) = c\int_a^1 da/(a^2H) \tag{1.15.5}$$

$$r = (1/\sqrt{-k})\sinh\left(\sqrt{-k}c\int_a^1 da/(a^2H)\right)$$

$$D_A = (1/\sqrt{-k})\sinh\left(\sqrt{-k}c\int_a^1 da/(a^2H)\right)$$

where, except for the first line, we have assumed $k < 0$. I leave it to the student to work out the $k > 0$ case. The $k = 0$ case should also be clear.

Box 1.15.3

Exercise 15.3.1: In calculating D_A vs. a, what are the two different ways curvature makes a difference?

Box 1.15.4

We have defined the density parameters $\Omega_x = \rho_{x,0}/\rho_{c,0}$ where $\rho_c \equiv 3H^2/(8\pi G)$ is the critical density, defined to be the total density for which the curvature, k is zero.

Exercise 15.4.1: Using the Friedmann equation, convince yourself that if $\rho = \rho_c$, then $k = 0$.

With this notation we can write

$$H^2(a) = H_0^2\left(\Omega_\Lambda + \Omega_m a^{-3} + \Omega_K a^{-2}\right) \tag{1.15.6}$$

where $\Omega_K \equiv -k/(H_0^2)$.

Exercise 15.4.2: Show that Eq. 1.15.6 can be derived from the Friedmann equation and the fact that $\rho_\Lambda \propto a^0$ and $\rho_m \propto a^{-3}$.

Exercise 15.4.3: Further, show that $\Omega_\Lambda + \Omega_m + \Omega_K = 1$.

Apparent and Absolute Magnitudes and the Distance Modulus

Magnitudes are absurd but useful if you want to use data from astronomers. They are a means of expressing luminosity and flux.

Luminosity: The luminosity of an object, L, is its power output. Usually its total electromagnetic power output, sometimes referred to as *bolometric* luminosity. Typical units for luminosity are ergs/sec (10^7 erg = 1 Joule, 1 Watt = 1 Joule/sec) or solar luminosity, $L_{\rm Sun}$. The Sun, by definition has a luminosity of one solar luminosity and $L_{\rm Sun} = 3.826 \times 10^{33}$ erg/sec = more than 10^{24} 100 Watt light bulbs. (You can remember this if you remember it's about as luminous as 7 Avogadro's number of 100 Watt light bulbs).

Flux: The flux, F, from an object is not an intrinsic property of the object, but also depends on the distance to the object. It is the amount of energy passing through a unit area, per unit of time. For an isotropic emitter in a non-expanding, Euclidean three-dimensional space, $F = L/(4\pi d^2)$ where d is the distance between source and observer. This equation just follows from energy conservation; note that the total power flowing through a spherical shell of radius d completely surrounding the emitter at its center is $4\pi d^2 \times F = L$.

Spectral Flux density: We usually do not measure the total flux from an object, but instead measure the flux in a manner that depends on how the flux is spread out in frequency. Thus a useful concept is the spectral flux density, S, that quantifies how much flux there is per unit frequency. The units of spectral flux density are erg/s/m^2/Hz. Where Hz is the unit of frequency called Hertz, equal to 1/s.

Apparent magnitude: Astronomers often use apparent magnitude, m, instead of flux. The apparent magnitude has a logarithmic dependence on flux; the reason for this is historical, and is fundamentally due to the logarithmic sensitivity of our eyes to flux. Not only is it logarithmic instead of linear, but *brighter* objects have *smaller* magnitudes. This is because the Greeks defined the brightest stars as stars of the first magnitude, and next brightest as stars of the 2nd magnitude, down to the stars we could just barely see at all, which are stars of the 6th magnitude. This ancient system, updated with precise definitions related to flux is still in use today (otherwise I would not bother telling you about it). One way of relating apparent magnitude to flux is the following:

$$m = M_{\rm Sun} - 2.5\log_{10}\left(\frac{F}{F_{\rm Sun10}}\right) \tag{1.15.7}$$

where $M_{\text{Sun}} = 4.76$ is the absolute magnitude of the Sun (see next definition) and F_{Sun10} is the flux we would get from the Sun if it were 10pc away. Since $L_{\text{Sun}} = 3.826 \times 10^{33}$ erg/sec and 1pc = 3.0856×10^{18} cm we get $F_{\text{Sun10}} = 3.198 \times 10^{-7}$ erg/cm^2/sec.

Note that because of the -2.5 factor in front of the $\log_{10}$, if the flux increases by a factor of 10, the apparent magnitude decreases by -2.5. Conversely, if the magnitude increases by 1, the flux decreases by a factor of $10^{1/2.5} = 10^{0.4}$.

Absolute Magnitude: The absolute magnitude of an object, denoted by M, is another way of expressing its luminosity. One way of defining it is via:

$$M = M_{\text{Sun}} - 2.5 \log_{10} \left(\frac{L}{L_{\text{Sun}}} \right). \tag{1.15.8}$$

Putting this together with the apparent magnitude-flux relationship above, one can show that this means $M = m$ for an object at 10pc.

Distance Modulus: The distance modulus is defined as $\mu \equiv m - M$. Note that as a difference between apparent and absolute magnitudes, this is equal to a log of the ratio of flux and luminosity. By plugging in the definitions above of m and M one finds

$$\mu = 5 \log_{10} \left(\frac{d_L}{10\text{pc}} \right) = 5 \log_{10} \left(\frac{d_L}{1\text{pc}} \right) - 5 \tag{1.15.9}$$

This page titled 1.15: Distance and Magnitude is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.16: Parallax, Cepheid Variables, Supernovae, and Distance Measurement

We have seen from the previous chapters, at least on very large scales, the Universe is the same everywhere and that it is expanding. Key to observing the consequences of this expansion is the ability to measure distances to things that are very far away. Here we cover the basics of how that is done. We have to do it in steps, getting distances to nearby objects and then using those objects to calibrate other objects that can be used to get to even further distances. We refer to this sequence of distance determinations as the distance ladder.

The first rung on this ladder is the use of trigonometric parallax to determine distances to the nearest stars. Some of these nearest stars are Cepheid variable stars with a luminosity that varies over time in a periodic manner. These stars have a relationship between the period of their luminosity variation and average lumniosity. With distances determined to some of them, that relationship can be calibrated. Once calibrated, then by determining their period, we can determine their luminosity and use them as standard candles to measure even further distances. Cepheids, in turn, can be used to calibrate type Ia supernova explosions. Supernovae are much much brighter than Cepheids, allowing us to observe them to even greater distances.

There's a very good and entertaining 3Blue1Brown YouTube video about the distance ladder you might want to check out. The video covers parallax and Cepheid variable stars, providing more information about these steps in the distance ladder than we manage to include here, and with some pretty cool animations. It also covers an earlier rung in the distance ladder we entirely neglect here: how to determine the Earth-Sun distance. Unfortunately it neglects supernovae, and treats redshift measurements as if they are a means to get a distance measurement -- which they are if you assume a specific cosmological model. This is a different perspective from our own, since we want to use measurements of distance and redshift to determine cosmological model parameters. Thanks to UC Davis student Peyton Harris for pointing this video out to me.

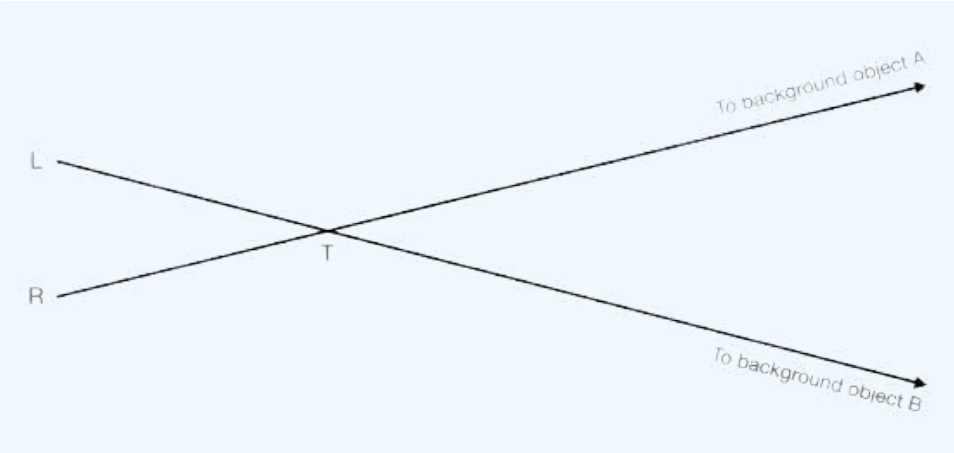
Direct observations - Parallax

Parallax is the shift in apparent location of a nearby object relative to more distant objects, as one changes the observation point. The phenomenon can readily be observed by holding your thumb out in front of you and switching your viewpoint from left eye to right eye and back. Simple trigonometry, and the small-angle approximation, lead to the relationship $d = \text{baseline}/p$ where p is the "parallax angle" in radians, and the baseline is half the distance between your two eyes.

The drawing below in the astronomical context has the parallax angle labeled on an angle with vertex at the nearby star. This is NOT actually the angle that gets directly measured, but rather is geometrically inferred from an angle that is measured.

Box 1.16.1

Exercise 16.1.0: In the figure below, L is left eye, R is right eye, and T is thumb held out at arm's length. Assume that the background objects that line up with the eyes and the thumb (both A and B) are effectively infinitely far away. i) Draw the angle (by adding it to the figure below) that one can actually observe with your right eye. Note: you will need to add the line from R to background object B in order to indicate this angle. ii) Draw in the baseline as described above, b. iii) Using the small-angle approximation and assuming the measured angle is in radians, relate the thumb-face distance to b and the observed angle you indicated in part i.



Exercise 16.1.1: Use observations of trigonometric parallax to estimate the length of your arm in units of your inter-ocular distance (IOD). That is, how many times longer is your arm than the space between your eyes. Since you don't have a protractor on you to measure angles, I'll tell you that your thumb, at the joint closer to the tip, held at arm's length subtends an angle of about 2 degrees. What is, roughly, the distance between your eyes in cm? Does the result you get for the length of your arm make sense?

Answer

It's about 6 cm between one pupil and the other. I got about 3 thumb widths for the shift from viewing with one eye to the other. This is six degrees. So the distance is about 6cm / sin(six degrees) = 6cm/(6*3.14/180) = 60 cm where in the first equality we used the small angle approximation after converting from degrees to radians. (Note the possibility here for making some mistakes with factors of two. The angle we measure here is 2p, as p is defined below, but the distance we are using is twice the baseline, as the baseline distance is defined below. These factors of 2, at least in the small-angle approximation, cancel out.)

The change in a star's position in the sky as a result of its true motion through space is called proper motion. This is distinguished from the annual apparent motion in the sky caused by the Earth's orbit around the Sun. A nearby star's apparent movement against the background of more distant stars is referred to as stellar parallax.

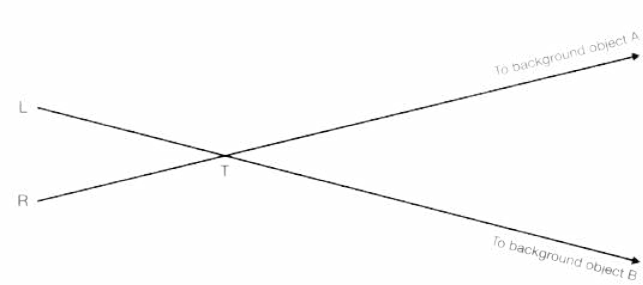
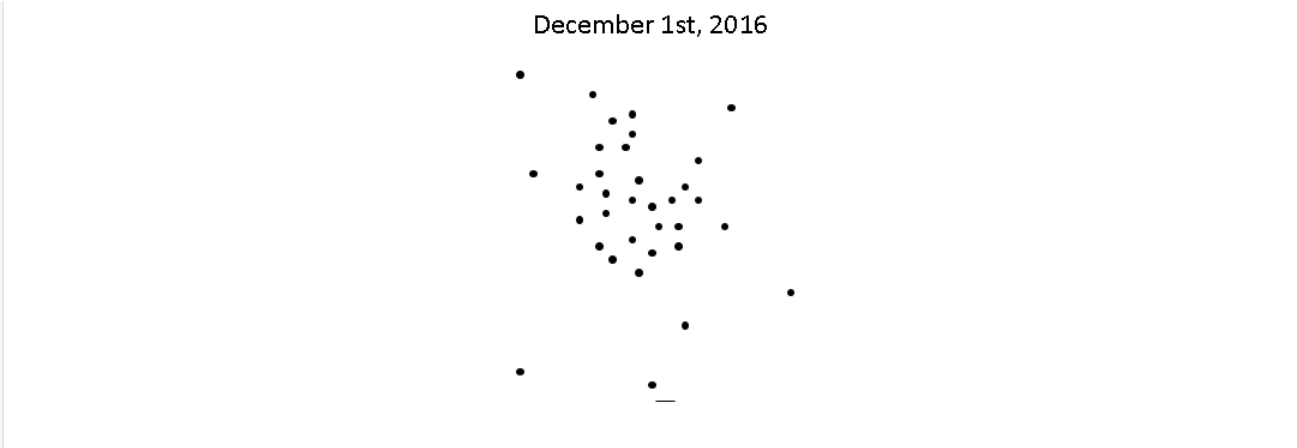


Figure 1.16.1: This exaggerated view shows how we can see the movement of nearby stars relative to the background of much more distant stars.Index{1}\): Copy and Paste Caption here. (Copyright; author via source)

Box 1.16.2



These images of the sky toward star cluster Knox0325 were taken exactly 6 months apart. Each dot is a star. The scale of angular separations is indicated by the line segment toward the bottom of the image which has a length of 0.03 arcseconds.

Exercise 16.2.1: Use trigonometric parallax to estimate the distance to star cluster KNOX0325 in units of the Earth-Sun distance, known as an astronomical unit or AU. There are 60 arc seconds in an arc minute and 60 arc minutes in one degree.

Answer

The observed shift is actually 2p, with p as indicated in the diagram. Relative to the background stars the cluster of stars is shifting by 0.03 arc seconds. To understand this, you have to realize the background stars are much, much further away than the Earth-Sun separation -- much further than it looks in the diagram. Because of this large distance, the line of sight from the December position to the blue background star is parallel to the line of sight from the June position to the blue background star. The angle at December formed by red background star -- December -- blue background star you cna then see is 2p. This is the observed shift. So p = 0.015 arc seconds. Looking at the geometry in the figure we have $d = 1AU/\sin(p) = 1AU/(\pi/(180*60*60)*0.015) = 1.4\times 10^7$AU.

Exercise 16.2.2: One AU is equal to 1.5×10^13 cm. How far away is KNOX0325 in cm? How far away is it in light years? A light year is the distance light travels in a year and the speed of light is 3×10^10 cm/sec.

Answer

Distance to KNOX0325 is 2.1×10^{20} cm = 220 light years.

Look back at the exaggerated stellar parallax image. The distance to the star is inversely proportional to the parallax. The distance to the star in parsecs is given by

$$d = \frac{1}{p}, \tag{1.16.1}$$

where p is in arc seconds.

The nearest star is proxima centauri, which exhibits a parallax of 0.762 arc sec, and therefore is 1.31 parsecs away.

Box 1.16.3

Exercise 16.3.1: The "parallax angle", p, is defined in such a way that the observed angular shift is equal to $2p$. A "parsec" is defined so that one parsec is the distance to an object with $p = 1$ arc sec. How many parsecs away is KNOX0325? The parsec (and kiloparsec, megaparsec and even gigaparsec) is a common unit of measure in cosmology. These are often abbreviated as pc, kpc, Mpc, Gpc.

Answer

d/parsec = 1 arscec/p ==> d = (1/.015) parsec = 66 parsec.

The limit of measurement from telescopes on the Earth's surface is about 20 parsecs, which only includes nearly 2000 of our closest stars. However, the distance at which parallax can be reliably measured has now been greatly extended by space-based instruments like the Hipparcos satellite and more recently the Gaia satellite.

Box 1.16.4

Exercise 16.4.1: The smallest angular separations that can be measured on the sky, so far, are 0.001 arc seconds (2025 update: about 10 times smaller now). To what distance can parallax be used for determining distances? You can give your answer in parsecs.

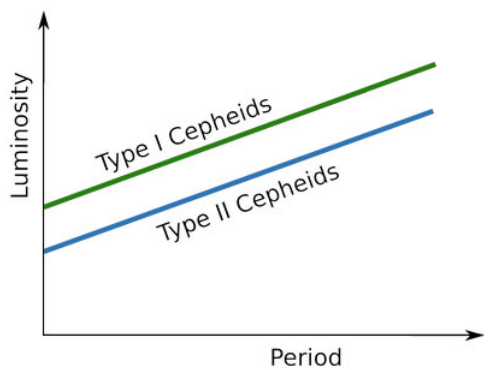
Answer

Answer: d/parsec = arcsec/p ==> d = 1,000 parsecs.

While parallax is used to calibrate the cosmic distance scale by allowing us to work out the distances to nearby stars, other methods must be used for much more distant objects, since their parallax angle is too small to measure accurately.

Standard Candles

While stellar parallax can only be used to measure distances to stars within hundreds of parsecs, Cepheid variable stars and supernovae can be used to measure larger distances such as the distances between galaxies and even galaxy clusters. Cepheid variable stars are intrinsic variables which pulsate in a predictable way. In addition, a Cepheid star's period (how often it pulsates) is directly related to its luminosity. The Hipparcos satellite mentioned earlier helped calibrate Cepheid distance scales by measuring the parallaxes of galactic Cepheids.



Cepheid variables are extremely luminous and very distant ones can be observed and measured. Once the period of a distant Cepheid has been measured, its luminosity can be determined from the known behavior of Cepheid variables. Then its absolute magnitude and apparent magnitude can be related by the distance modulus equation, and its distance can be determined.

$$d = 10^{(m-M+5)/5} \tag{1.16.2}$$

- d is the luminosity distance to the object in parsecs
- m is the apparent magnitude of the object
- M is the absolute magnitude of the object

Box 1.16.5

Exercise 16.5.1: There is a Cepheid in Galaxy A with a period of 30 days and an apparent magnitude of $m = 26$. How far away is Galaxy A? Use the fact that the Cepheid period-luminosity relation says that a Cepheid with a period of 30 days has an absolute magnitude of $M = 11$. (Note: the answer is 10kpc which is still within our own galaxy so it does not make any sense. I need to update this with a larger apparent magnitude so that it will be further away. If I made it m=36 that would make it 100 times further, which would make it 1Mpc which makes sense. I would then also need to update m for the supernova in Galaxy A (see next Exercise)).

Answer

$d = 10^{(26-11+5)/5}$ parsec = 10^4 parsec.

Cepheid variables can be used to measure distances from about 1 kpc to 50 Mpc. Beyond 50 Mpc it becomes too difficult to separate out the light that is just from a Cepheid, from the light from nearby stars. Astronomers call this problem "crowding."

Type Ia supernovae are all caused by exploding white dwarfs which have companion stars. The gravitational pull of the white dwarf causes it to take matter from its companion star. Eventually it reaches a high enough mass that it cannot support itself against gravitational collapse and explodes. All type Ia supernovae reach nearly the same brightness at the peak of their outburst. They then follow a distinct curve as they decrease in brightness. So when astronomers observe a type Ia supernova, they can measure its apparent magnitude at peak brightness. If they know the distance to the supernova, perhaps becaues they have determined using Cepheids the distance to the galaxy hosting the supernova, they can then determine the absolute magnitude of the supernova. We refer to this as supernova calibration. Now, if another supernova is observed, assuming it has the same peak brightness absolute magnitude, we can use measurement of its apparent magnitude at peak brightness to determine its distance.

Box 1.16.6

Exercise 16.6.1: There was also a supernova explosion in Galaxy A (from the previous problem) whose brightness varied with time, but at peak brightness had an apparent magnitude of $m = -4.3$. What is the absolute magnitude of the supernova at peak brightness?

Answer

We know the distance to Galaxy A is 10^4 parsecs. So we can use $d/parsec = 10^{(m-M+5)/5}$ to solve for M. Or, we could just say that since d is the same for the Cepheid and the supernovae, the m-M values have to be the same. For the Cepheid, m-M = 15, so for the supernova m-M=15. This implies that M=m-15 =-19.3.

Exercise 16.6.2: Another supernova went off in Galaxy B and had an apparent magnitude at peak brightness of $m = 15.7$. Assuming supernovae peak brightnesses are standard candles, how far away is Galaxy B?

Answer

$d = 10^{(15.7+19.3+5)/5}\text{parsec} = 10^8$parsec = 100 Megaparsec.

Type Ia supernovae can be distinguished from other supernovae because they do not have hydrogen lines in their spectra and have a strong Si II line at 615 nm. The peak of their outburst has an absolute magnitude of -19.3±0.03.Type Ia supernovae can be used to measure distances from about 1 Mpc to over 1000 Mpc.

A Standard Ruler

If you observe an object of known length, and determine the angle it subtends, then you can determine the angular diameter distance to the object. As described in the previous chapter, if x is the comoving length of the object and θ is the angle it subtends when oriented so that length runs perpendicular to the line of sight, then the comoving angular diameter distance is, in the small-angle approximation, $D_A = x/\theta$. In the previous chapter we also saw that $D_A = d_L/(1+z)$ where d_L is the luminosity distance.

A very important standard ruler in cosmology is the *sound horizon*. The sound horizon is the distance that a sound wave can travel through the plasma of the big bang, from the beginning until the plasma disappears. We usually denote the comoving sound horizon as r_s. Assuming the standard cosmological model, an estimate of the comoving sound horizon given data from the Planck satellite is $r_s = 147.09 \pm 0.26$ Mpc. The sound horizon leaves an imprint in the matter distribution. We can observe galaxies near some redshift z and measure the statistical properties of both their angular distribution and their distribution with redshift. From this we can infer how the sound horizon projects into an angle perpendicular to the line of sight, θ_s and how it projects into a redshift separation along the line of sight, δz_s. We thus get both a distance estimate: $D_A = r_s/\theta_s$ and an estimate of $H(z)$ since it turns out that $r_s = \delta z_s/H(z)$.

We can calculate r_s if we know the sound speed c_s and the expansion rate, $H(a)$, and the scale factor when the plasma disappears, a_d. In time dt the sound wave travels a comoving distance (physical distance divided by scale factor) of $c_s dt/a(t)$ so

$$r_s = \int_0^{a_d} da \frac{c_s(a)}{a^2 H(a)}. \tag{1.16.3}$$

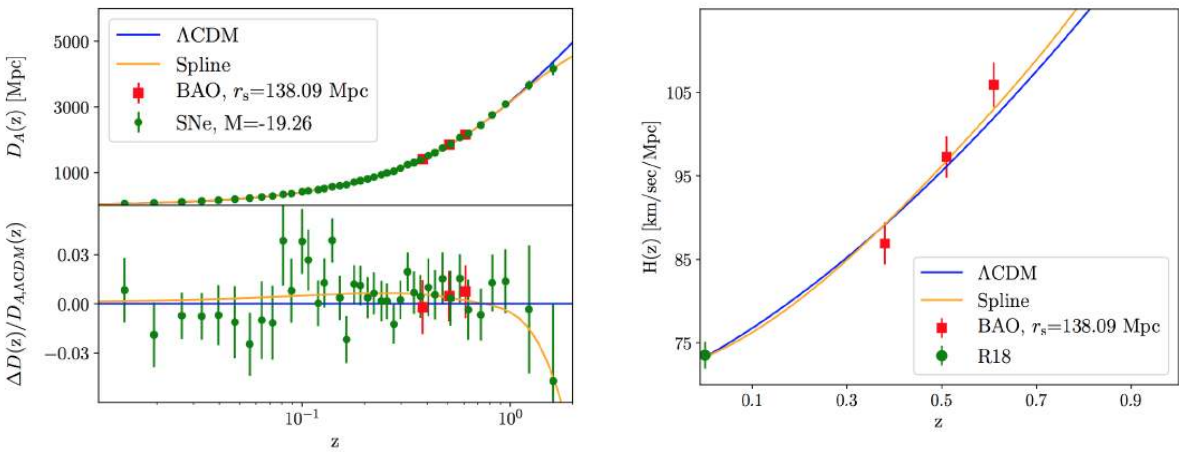


Figure 1.16.1: Figures are from Aylor et al. (2019). Left panel: Comoving angular diameter distance as a function of redshift. Green data points: inferences from the Scolnic et al. (2018) measurements of supernova apparent magnitude, assuming M = -19.26. Red data points: inferences of distances from galactic baryon acoustic oscillation (BAO) data given a comoving sound horizon of $r_s = 138.09$ Mpc. Model curves are for the ΛCDM model that best fits the Cepheids plus supernovae + BAO data, and the "Spline" model that does the same (as described in Aylor et al. (2019)). The botttom of the left panel shows residuals after subtraction of the ΛCDM model. Right panel: Hubble parameter as a function of redshift. Green data point is the Riess et al. (2018) result for H_0. Red data points are inferences of $H(z)$ from galactic BAO data given a comoving sound horizon of $r_s = 138.09$ Mpc.

HOMEWORK Problems

These are all to be done with a computer, except for the first one.

Problem 1.16.1

In the problem following this one you are going to want to have error bars for the distance measurements. But the measurements are actually reported as apparent magnitudes. So you will need to propagate magnitude errors to distance errors. Because a small change in distance, δD_A is related to a small change in apparent magnitude δm by $\delta D_A = (\partial D_A/\partial m)\delta m$, you can write $\sigma^2(D_A) = (\partial D_A/\partial m)^2 \sigma^2(m)$. Show that, as a result, $\sigma(D_A)/D_A = 0.2\ln(10)\sigma(m)$..

Problem 1.16.2

Plot up $D_A(z)$ vs. z for the supernova data assuming $M = -19.3$ which is about what the Cepheid calibration of supernovae give for their absolute magnitude at (corrected) peak brightness. Include error bars in your plot. Label the axes appropriately.

Problem 1.16.3

Calculate $D_A(z)$ vs. z for 4 theoretical models. The four cases to include are i) $H_0 = 73$ km/sec/Mpc; $\Omega_m = 0.3, \Omega_\Lambda = 0.7, \Omega_k = 0$, ii) $H_0 = 67$ km/sec/Mpc; $\Omega_m = 0.3, \Omega_\Lambda = 0.7, \Omega_k = 0$, iii) $H_0 = 73$ km/sec/Mpc; $\Omega_m = 1, \Omega_\Lambda = 0, \Omega_k = 0$, iv) $H_0 = 73$ km/sec/Mpc; $\Omega_m = 0.3, \Omega_\Lambda = 0, \Omega_k = 0.7$. Plot up the $D_A(z)$ curves with two different z ranges: i) enough to cover all the data and ii) over the interval 0 to 0.2. For both of these choose an appropriate y axis range. Include the data, with error bars, from 16.2 in your plots. This should just be 2 plots.

Problem 1.16.4

Answer these questions based on the graphs in the above problem. Is the z < 0.2 redshift interval relatively insensitive to the density parameters? What parameter is this lower-redshift data sensitive to? Over the whole redshift range, which model would you say provides the best fit to the data? Of the two Hubble constants given, which provides a better fit to the data?

Problem 1.16.5

We use the statistical quantity χ^2 as a measure of the quality of agreement of a model prediction with the data. Usually, the lower χ^2, the better the agreement. Assuming $M = -19.3$, calculate χ^2 for the 4 above models where

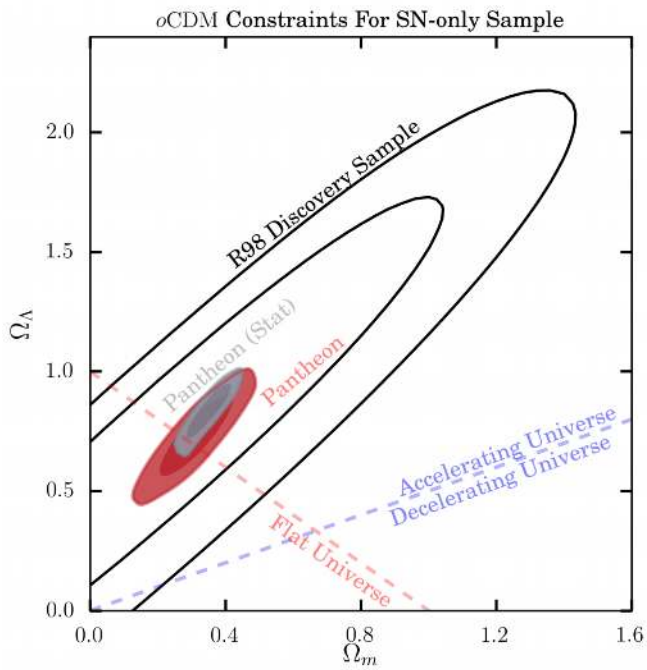
$$\chi^2 = \Sigma_i \left(m_i^d - m_i^m\right)^2 / \sigma_i^2 \tag{1.16.4}$$

with m_i^d the measured apparent magnitude of the ith supernova (with 'd' for 'data'), m_i^m is the apparent magnitude of the ith supernva as predicted by the model, and σ_i is the error on the ith magnitude measurement.

This page titled 1.16: Parallax, Cepheid Variables, Supernovae, and Distance Measurement is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

- **Current page** by Lloyd Knox is licensed CC BY 4.0.
- **2.15: S17. Parallax, Cepheid Variables, Supernovae, and Distance Measurement - SOLUTIONS** by Lloyd Knox is licensed CC BY 4.0.

1.17: Cosmological Data Analysis



We focus in this chapter on the analysis of cosmological data. Most of what I present in this chapter applies much more broadly; in fact, nothing in this text book is of broader utility. However, the presentation here is entirely focused on application to cosmology. We wish to learn from measurements of the cosmos. These measurements are never 100% precise, and thus we need a means of dealing with uncertainty. We necessarily deal in probabilities.

This is especially true for the practitioner interested in discovering something new. Almost always, the data are not overwhelmingly and obviously convincing of the new truth that they potentially reveal. Data analysis in cosmology is an exciting process of sorting out whether one is on the cusp of an exciting discovery, or on the cusp of embarrassing oneself by claiming something that turns out not to be true. It calls for rigor of process and high integrity. Engaging in it, in the right way, will raise one's standards of what it means to know something. One needs to search not only for the evidence that supports one's hunch of what's going on, but also, very importantly, one needs to search for evidence that supports alternative explanations. We are after the truth.

There are two distinct aspects of data analysis: model comparison and parameter estimation. In model comparison we try to determine which model is better than another. In parameter estimation, we have one assumed model and we are estimating the parameters of that model. We focus here on parameter estimation.

Modeling data: a simple example

Let's start by considering a simple measurement of length, to have a specific example in mind. The signal would be the true length, ℓ, and our model of the data might be

$$d = \ell + n \tag{1.17.1}$$

where d is the measurement (our data) and n is the error in the measurement, the difference between the data and the true length, ℓ. We are fundamentally interested in the probability distribution of the length, given this data. The length is the one parameter of our model. (More generally, we are interested in the joint probability distribution, given the data, of all the parameters of our model.)

A measurement, if it is to mean anything at all, has to come along with an estimate of the uncertainty in the measurement. How to estimate the uncertainty is a subject we won't explore here. We are going to assume the measurement has been done and the uncertainty in it has been accurately determined. We will further assume here for simplicity that the uncertainty can be described with a normal distribution. In our one-dimensional example this means that

$$P(n) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{n^2}{2\sigma^2}) \tag{1.17.2}$$

is the probability density for the error, n. What is a probability density you ask? This one tells us the amount of probability that there is between n and $n+dn$: it's equal to $P(n)dn$. The pre-factor out in front of the exponential is there to keep $P(n)$ properly normalized so that

$$\int_{-\infty}^{\infty} dnP(n) = 1. \tag{1.17.3}$$

It should integrate to 1 because n takes on one and only one value (even if we don't know what that value is).

A fundamentally important quantity is the probability of some data, d, given the signal s (or, in our example, ℓ). Since $n = d - \ell$ we can write

$$P(d|\ell) = P(n) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(d-\ell)^2}{2\sigma^2}\right) \tag{1.17.4}$$

as the probability density of d given that the length is ℓ. That is, if we knew the true length was ℓ then how probable is it to have d in the range $[d, d+dd]$? Answer: $P(d|\ell)dd$. (What we mean here by dd is an infinitesimal increment to d).

Bayes' Theorem

Although we have an expression that tells us the probability of the data given the true value of the underlying parameter, we are actually in a position of wanting the exact opposite! We know what the data is and we desire to know what the true length is -- or, more preciesly, since we can't have perfect knowledge of the length, our goal is to know the probability that the length takes on any particular value: we want $P(\ell|d)$. Bayes' theorem helps us get from one to the other. It follows from fundamental axioms of probability theory. For simplicity, for the purposes of deriving Bayes' theorem, we are going to start considering discrete outcomes, like we get with the flip of a coin, or the roll of dice. Let's introduce the joint probability $P(A, B$ where A might be a particular outcome for a six-sided die and B might be the sum of that outcome with that of another die. I've purposely constructed this example so that A and B are not independent. By the joint distribution, we mean that $P(A, B)$ tells us the probability that the first quantity is A and the second quantity is B.

I'll give you the calculation of $P(A = 3, B = 8)$. There's only one way for this to happen. The first die turns up 3 and the second one turns up 5. With a roll of two die this is one of 36 possible and equivalent outcomes, so the probability is 1/36. Of course, the probability that $B < A$ is zero.

Exercise: Calculate these probabilities given the above definition of A and B: $P(A = 5, B = 6)$, $P(A = 5, B = 3)$,$P(A = 3, B = 7)$.

It is a fundamental rule of probability that $P(A, B) = P(A|B)P(B)$. The joint distribution is symmetric so it is also true that $P(A, B)$ $= P(B|A)P(A)$.

Exercise: From the preceding two equations derive $P(A|B) = P(B|A)P(A)/P(B)$.

In our special case of interest, if we call our model parameters θ this becomes Bayes' theorem:

$$P(\theta|d) = \frac{P(d|\theta)P(\theta)}{P(d)}. \tag{1.17.5}$$

Bayes' theorem describes learning from data. When we learn something, we're usually not starting from complete ignorance. The factor $P(\theta)$ is called the prior probability distribution, or simply, 'the prior.' It represents what we know about the model parameters prior to examining the data. The probability density $P(d|\theta)$ (at least when thought of as a function of \theta) is called the likelhood. The denominator, $P(d)$, I find difficult to understand conceptually. It's the probability of the data, but we know the data. That's confusing. But there is an easier way to think about it for purposes of parameter estimation: it is just a normalizing constant. The reason we can do this is that, by assumption, the model is correct and so θ must take on one value; i.e., if we integrate the posterior over all possible values of θ then it must be equal to 1. Forcing this to be true will determine the value of $P(d)$ if one knows the likelihood and the prior. One can do the integral without knowing $P(d)$ because $P(d)$ does not depend on θ.

The term on the left-hand side of the above equation is called the 'posterior probability distribution', or sometimes simply 'the posterior.' Getting back to thinking of Bayes' theorem as a description of learning from data, we can now see that the likelihood serves to update our prior beliefs, incorporating what we've learned from data, and what we know already, into the posterior probability distribution for θ; i.e., what we know about θ after we have studied the data. Usually we don't care about the normalization -- we just want to know how probable one value of θ is compared to another.

Modeling measurements of supernova apparent magnitude

Let's now turn to the case of modeling our supernova data. Let's take the model parameters to be $\theta = \{M, \Omega_\Lambda, \Omega_m, H_0\}$. The data we take to be the supernova apparent magnitudes. The likelihood we take to be normally distributed (as we are assuming the errors in the magnitudes are normally distributed). Therefore we have

$$L(\theta) = P(d|\theta) \propto \exp(-\chi^2/2) \;\; ; \;\; \chi^2 = \Sigma_i (m_i^d - m_i^m)^2/\sigma_i^2 \tag{1.17.6}$$

where

$$m_i^m = M + 5\log_{10}\left(\frac{D_L(H_0, \Omega_\Lambda, \Omega_m; z_i)}{\text{Mpc}}\right) + 25. \tag{1.17.7}$$

We take a prior that is uniform in Ω_Λ, Ω_m and H_0 and normally distributed in M so that

$$P(\theta) \propto \exp(-\chi_M^2/2) \;\; ; \;\; \chi_M^2 = (M - (-19.26))^2/\sigma_M^2 \tag{1.17.8}$$

with $\sigma_M = 0.05$. This mean and σ for M is consistent with the Riess et al. (2018) determination of the Hubble constant with a standard error of 2.2%, assuming that uncertainty is entirely due to uncertainty in supernova absolute magnitude calibration. We can now take the above prior and the above likelihood and multiply them together to form the posterior (up to an unknown normalization constant that we don't care about).

Marginalization

It is often the case that we do not care about all the parameters of our model. Maybe we only care about Ω_m and Ω_Λ. In that case we might want to calculate $P(\Omega_m, \Omega_\Lambda|d)$. This probability density should include the probability associated with all values of the other parameters. It is related to the full joint distribution by integration over the other parameters via:

$$P(\Omega_m, \Omega_\Lambda|d) = \int dH_0 dM P(H_0, \Omega_\Lambda, \Omega_\Lambda, M|d). \tag{1.17.9}$$

This process of integrating over parameters is called marginalization.

Contour plots

A common way of presenting a two-dimensional probability distribution is a contour plot. The axes of the contour plot are the two parameters, and the contours indicate curves of constant probability density. Often there will be two different curves plotted: one that encloses 68% of the probability and another that encloses 95% of the probability. The enclosed regions are the 68\% confidence region and the 95% confidence region, respectively.

An example contour plot is shown in the figure at the beginning of this chapter. The contours labeled "R98" discovery sample give the 68% and 95% confidence regions given the Riess et al. (1998) data that were used for the discovery of cosmic acceleration. The red and dark red shaded regions are the 68% and 95% confidence regions, respectively, given the Scolnic et al. (2018) supernova data, but only taking into account some of the sources of error. The authors distinguish some of their errors as systematic, as opposed to statistical. Including the systematic errors as well leads to the grey contours. The shrinkage from the R98 contours to the gray contours indicates the progress that has occurred in supernova cosmology over the 20 years from 1998 to 2018.

If the probability density is Gaussian then the contours are such that $2\ln(P_{peak}/P_{68}) = 2.3$ and $2\ln(P_{peak}/P_{95}) = 6.17$, where P_{peak} is the probability density evaluated at its maximum and P_{68} (P_{95}) is the probability density whose contour contains 68.3% (95.4%) of the data. (You might wonder where these extra significant figures come from to make this 68.3 and 95.4. They come from one-dimensional normal probability distributions. For a normal one-dimensional distribution, 68.3% of the probability is to be found in between one standard deviation (σ) less than the mean, and one standard deviation more than the mean while 95.4\% of the probability is to be found in between two standard deviations less than the mean, and two standard deviations more than the mean.)

Homework:

18.1: Estimate H_0. To reduce dimensionality, in order to make things simpler, set $\Omega_\Lambda = 1 - \Omega_m = 0.7$. Take M to be governed by the above prior. Use the Scolnic et al. (2018) data (available in 'supernova_data.txt') to make a contour plot in the H_0, M plane.

18.2: Starting from the above $P(H_0, M|d)$, approximate the integral $P(H_0|d) = \int_{-\infty}^{\infty} dM P(H_0, M|d)$ with a discrete sum and produce a plot of H_0 vs. $P(H_0|d)$. Don't worry about the normalization of the probability density you plot.

18.3: Produce your own Ω_m, Ω_Λ 68\% and 95\% confidence contours using the Scolnic et al. (2018) data and, for simplicity, fixing M=-19.26 and $H_0 = 72.9$ km/sec/Mpc. Ideally you would marginalize over these other variables, instead of fixing them, but that would be significantly more challenging. You can approximate the distributions as Gaussian (normal) for purposes of choosing the contour levels.

This page titled 1.17: Cosmological Data Analysis is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.18: The Early Universe

The hot big bang contained a hot, dense "soup" of elementary particles. Fermilab is a U.S. national laboratory in Illinois that is the home of the Tevatron, a particle accelerator that smashed protons into anti-protons at high energies to create new particles.

In the next chapters we study the "primordial soup" of the big bang. One of the first human beings to travel this intellectual territory was George Gamow born in 1904 in Russia as Georgi Antonovich Gamow. While at university of Leningrad, he studied under Friedmann (of Friedmann-Robertson-Walker fame) until Friedmann suddenly died in 1925, forcing Gamow to change dissertation advisers. Gamow's early accomplishments include in 1928 solving the problem of alpha decay of nuclei through quantum tunneling, for which the Gamow Factor (a probability factor involved in tunneling through the Coulomb barrier) bears his name. As the Soviet Union grew more oppressive, Gamow sought to escape to Europe but was denied permission to leave. Eventually Gamow and his wife managed to flee to the United States where he became a professor at George Washington University in 1934.

At GWU, Gamow turned his interest toward astrophysics and fathered the idea of the hot Big Bang. His main assumption was that the universe started at near-infinite temperature and density from which it rapidly expanded and cooled. In this early stage, the universe would consist of protons, neutrons, and electrons zipping around in a hot bath of electromagnetic radiation. He argued that one could build all the elements from the ground up via neutron capture, but in the end was never able to account for the abundances of elements higher than helium. Gamow joked that his theory was basically correct, since it accounted for 99% of the universe's rest mass in hydrogen and helium. Gamow published the paper, titled ``The Origin of Chemical Elements'' with his PhD student Ralph Alpher in the April 1948 issue of Physical Review. Gamow included famous contemporary astrophysicist Hans Bethe in the author list (without his notification, nevermind permission!) so that it would read like α, β, γ.

Alpher and Gamow found that the big bang had to be hot in order to avoid over production of the elements. A cold big bang would expand more slowly, allowing more time for build-up of heavy elements via neutron capture. Alpher and another researcher, Robert Herman, predicted that this heat should still be present today, in the form of black-body radiation cooled down by the expansion from billions of degrees to just about 5° Kelvin. This prediction was eventually verified with an accidental discovery in 1964 by Bell Labs astronomers Arno Penzias and Robert Wilson of an isotropic background with a temperature of about 3° K, a discovery that led to the establishment of the hot big bang as the dominant cosmological paradigm, as well as the 1979 Nobel Prize in physics for Penzias and Wilson.

It is now quite well-established that the universe used to be very hot, very dense, and expanding very rapidly. Relics that remain today from that era include most of the Hydrogen and Helium in the universe, trace amounts of Deuterium, Helium-3, and the photons that comprise the cosmic microwave background discovered by Penzias and Wilson. There is indirect evidence, via multiple channels, that there is also a cosmic neutrino background today, left over from the big bang. More speculatively, the dark matter is another relic of the big bang.

To understand this primordial soup and its relics, we now turn our attention from a relativistic understanding of the expansion of space, to statistical mechanics. We begin with equilibrium statistical mechanics, before moving on to a discussion of departures

from equilibrium. We will come to understand the production in the big bang of Helium, photons, other "hot" relics such as neutrinos, and "cold" relics such as the dark matter. We will also discuss the observations that test our ideas and constrain parameters of the standard cosmological model and its extensions.

This page titled 1.18: The Early Universe is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.19: Equilibrium Statistical Mechanics

Out of the early Universe we get the light elements, a lot of photons and, as it turns out, a bunch of neutrinos and other relics of our hot past as well. To understand the production of these particles we now turn to the subject of Equilibrium Statistical Mechanics.

Phase Space

A collection of particles is conveniently described by how it is distributed in both position and momentum. We usually assume three spatial dimensions, and in this case, the momentum of a particle is a three-dimensional quantity (it takes 3 numbers to specify the momentum of a particle, p_x, p_y and p_z). We refer to the space itself as "configuration space" (x, y, z) and the three-dimensional space associated with momentum as "momentum space." We can put these spaces together into one six-dimensional object (x, y, z, p_x, p_y, p_z) we call "phase space."

The phase space distribution function

We conventionally describe the location in phase space of large numbers of particles in a statistical manner, where we just state the average number of particles as a function of location in phase space. More specifically, we define a phase space distribution function, f, such that the number of particles at (x, y, z, p_x, p_y, p_z) in a phase-space volume of size $dx\,dy\,dz\,dp_x\,dp_y\,dp_z$ is

$$dN = \frac{f(\vec{x},\vec{p})}{h^3} dx\,dy\,dz\,dp_x\,dp_y\,dp_z \tag{1.19.1}$$

where h is Planck's constant. This equation serves to define f: It tells us the number of particles in a phase space volume equal to h^3.

Types of Equilibria

We are going to introduce some results of statistical mechanics that are quite amazing and useful and that apply in *equilibrium*. So let us first define equilibrium. In fact, we will define two different kinds of equilibrium: kinetic and chemical.

Kinetic equilibrium obtains when reactions that exchange energy between particles (such as collisions) occur rapidly compared to the time-scale under which conditions are changing. For example, the gas particles in this room interact very rapidly. For a given particle, the typical time between collisions is well below a second. Given that conditions in the room are not changing rapidly (that is, the temperature in the room is quite stable), the gas particles in the room will be in kinetic equilibrium.

Chemical equilibrium obtains when reactions that exchange particle type are rapidly occurring. An example of a reaction that changes particle type is electron, positron annihilation:

$$e^- + e^+ \rightarrow 2\gamma \tag{1.19.2}$$

where γ is a photon. Another example is given by chemical processes, and is where "chemical equilibrium" gets its name, such as

$$2\,\mathrm{H} + \mathrm{O} \rightleftharpoons \mathrm{H_2O}. \tag{1.19.3}$$

When these reactions are fast, chemical equilibrium is rapidly achieved. In chemical equilibrium, just as many forward reactions as backward reactions are happening (so the number densities of all the particles are independent of time).

Equilibrium forms for f

The first of two amazing results from statistical mechanics we will use (without proof) is the following. In *kinetic* equilibrium the phase space distribution function always has the following form:

$$f = \left[\exp\left(\frac{E(p)-\mu}{k_B T}\right) \pm 1\right]^{-1}. \tag{1.19.4}$$

where the + is for fermions and the - is for bosons, T is the temperature, μ is the chemical potential and E is the energy of each particle, $E^2 = p^2c^2 + m^2c^4$.

You already have some intuition for what is meant by temperature. We will soon do some exercises to see how the way that T affects f is consistent with your ideas about temperature. You may have less of a feeling about the physical meaning of μ. We'll do some exercises to address that. In the meantime, I'll tell you that when two systems are allowed to exchange kinetic energy, after a sufficiently long time if conditions are not changing, then their temperatures become related (in fact, equal). Similarly, if particles

are allowed to change *type*, then the chemical potentials of the different types of particles also become related (though not necessarily equal). For example, if this reaction is proceeding rapidly, in both the forwards and backwards direction:

$$a + b \longleftrightarrow c + d \tag{1.19.5}$$

then *chemical equilibrium* will obtain and the number densities n_a, n_b, n_c and n_d will be independent of time. Further, and this is our second result from statistical mechanics, we will have this relation between their chemical potentials:

$$\mu_a + \mu_b = \mu_c + \mu_d. \tag{1.19.6}$$

More generally, for

$$a + b + ... \longleftrightarrow c + d + ... \tag{1.19.7}$$

happening rapidly, then $\mu_a + \mu_b + ... = \mu_c + \mu_d + ...$

As a further example, if the reactions

$$e^- + e^+ \longleftrightarrow 2\gamma \tag{1.19.8}$$

are happening rapidly then $\mu_{e^-} + \mu_{e^+} = 2\mu_\gamma$.

How to go from f to number density, energy density and pressure

The number density, energy density and pressure of a collection of free particles is:

$$\begin{aligned} n(\vec{x}) &= g \int \frac{d^3p}{h^3} f(\vec{x}, \vec{p}) \\ \epsilon(\vec{x}) &= g \int \frac{d^3p}{h^3} E(p) f(\vec{x}, \vec{p}) \\ P(\vec{x}) &= g \int \frac{d^3p}{h^3} \frac{p^2c^2}{3E} f(\vec{x}, \vec{p}) \end{aligned} \tag{1.19.9}$$

where f is the phase-space distribution function and g counts the number of internal degrees of freedom for the particles. For example, electrons have two spin states so for them $g = 2$.

Box 1.19.1

Exercise 19.1.1: From the definition of f in Eq.1.19.1, derive the above expression for the number density $n(\vec{x})$ assuming that f is the same for each internal degree of freedom. (If this "internal degrees of freedom" part is concerning you, it's OK to simplify the problem a bit by doing it for a particle with a single internal degree of freedom, so $g = 1$; i.e., just ignore g.)

Answer

Since f/h^3 is the density of particles in phase space, the number in some small volume V, small enough such that f does not change much throughout the volume, is given by N = V$\int d^3pf/h^3$. So the number density is $n = \int d^3pf/h^3$. If every internal degree of freedom has the same value of f and if we want to count all the particles, regardless of internal state, then we have $n = g \int d^3pf/h^3$.

Exercise 19.1.2: From the definition of f, derive the above expression for the energy density $\epsilon(\vec{x})$.

Answer

Same as above except instead of calculating the total number in some small volume, we want the total energy. Therefore we just insert E(p) into the integral, to add up the energy from each region of momentum space, instead of just the number of particles from each region of momentum space.

Exercise 19.1.3: (Optional) From the definition of f, derive the above expression for the pressure $P(\vec{x})$. This one is significantly harder. You need to recall that pressure is force per unit area. The force on a wall from particles hitting it in time interval Δt is equal to the sum of the changes in each particle's momentum as it bounces off the wall, divided by Δt. First show that for a wall perpendicular to the x axis this force is given by $F = \frac{1}{2} \int d^3pfA(v_x\Delta t)(2p_x)/\Delta t$ so that $P = \int d^3pfv_xp_x$. Then use the fact that $v_x = p_xc^2/E$ (see footnote[2]) and that $\int d^3pfp_x^2 = 1/3 \int d^3pfp^2$ as long as f only depends on $p^2 = p_x^2 + p_y^2 + p_z^2$ to get P

$= \int d^3pfp^2c^2/(3E)$. If there are g internal degrees of freedom, that's that many more particles doing exactly the same thing (f tells us the distribution for each internal degree of freedom) so we get the desired result including the factor of g.

Answer

I have not produced a written solution for this yet.

An example: black body (thermal) radiation

From Equation 1.19.4 and Equation 1.19.9 we can derive a lot of results. For example, a gas of photons ($g = 2$) in kinetic equilibrium with $\mu = 0$ (how this arises physically to be explained later) has a contribution to its number density from particles with magnitude of momentum between p and $p + dp$ equal to

$$n(p)dp = \frac{8\pi}{h^3}\frac{p^2dp}{\exp(pc/(k_BT)) - 1}. \tag{1.19.10}$$

Note that by $n(p)$ we mean the function of the magnitude of momentum that when integrated over p gives number density $n = \int_0^\infty dp$ $n(p)$.

Box 1.19.2

Exercise 19.2.1: Derive the above equation for $n(p)dp$. Start from the integral above (one of equations 1.19.9) that gives the number density. Because the energy of each particle (and therefore the whole integrand) only depends on the magnitude of the momentum, $p = \sqrt{p_x^2 + p_y^2 + p_z^2}$, switch from Cartesian to spherical coordinates and integrate over the angular variables. That is, replace $d^3p = dp_xdp_ydp_z$ with $p^2dpd(\cos\theta_p)d\phi_p$ and integrate over the angular variables θ_p and ϕ_p. Remember that $E(p) = pc$ for photons.

Answer

The integral over angular variables results in $n = \frac{g}{h^3}\int_0^\infty 4\pi p^2dpf$. Now the integral is adding up numbers of particles from momentum space shells of momentum space volume $4\pi p^2dp$. We need only substitute in the appropriate expression for f, set $E = pc, \mu = 0$ and $g = 2$ to get the answer.

Exercise 19.2.2: If you sampled one photon out of the distribution, there is a probability that it will have a magnitude of momentum between p and $p + dp$. For what value of p does this probability peak? Note that answering this question will require you to solve a transcendental equation. That's a bit too much work so instead you can just show that $p = 1.58k_BT/c$ is quite close to the peak.

Answer

To find the peak of the distribution set $dn(p)/dp = 0$ and solve for p. One ends up with a transcendental equation to solve. Setting $x = pc/kT$ makes it possible to write it down fairly compactly as $(2 - x)e^x = 2$. One can narrow in on the solution numerically with a calculator -- especially a graphing one if you just plot the left-hand side and choose the value of x that gives 2. I used a calculator and in a few tries had $x = 1.6$ which is pretty close as $(2 - 1.6)e^{1.6} = 1.98$. So the most probable p is $p = 1.6k_BT/c$.

Exercise 19.2.3: How does that most probable p depend on temperature? Notice that this is qualitatively consistent with what you expect for temperature.

Answer

The most probable p corresponds to an energy that is $pc = 1.6k_BT$. This corresponds to what we expect since we see that $\simeq k_BT$ is a typical particle kinetic energy.

To perform the integral over p and obtain an expression for the number density of photons in kinetic equilibrium with zero chemical potential, we make a change of variables $x = pc/(k_BT)$ to remove all dimensionful constants from the integral and find:

$$\int_0^\infty n(p)dp = \frac{8\pi}{c^3h^3}(k_BT)^3\int_0^\infty \frac{x^2dx}{\exp(x) - 1}. \tag{1.19.11}$$

The integral can be looked up in a table or performed numerically. It's equal to $2\zeta(3) \simeq 2.404$ where ζ is the Riemann zeta function.

What physical conditions lead to $\mu = 0$?

If a particle can be freely created or destroyed, without other particles being created or destroyed, and these reactions are sufficiently fast, then the chemical potential will be driven towards zero. We can see this from the rule we already learned.

Assume this reaction is fast, called free-free, or Bremstrahlung:

$$e^- + p^+ \rightarrow e^- + p^+ + \gamma \tag{1.19.12}$$

in which an electron is accelerated in the electric field of a proton and thus radiates a photon. If the photon can get absorbed in some way, then we also effectively have the reverse reaction as well. In this case we would have $\mu_{e^-} + \mu_{p^+} = \mu_{e^-} + \mu_{p^+} + \mu_\gamma$ which leads us to $\mu_\gamma = 0$.

Equilibrium Statistical Mechanics Results in Various Limits

All of these results come from doing the appropriate integral over $f = (\exp[(E(p) - \mu)/(k_B T)] \pm 1)^{-1}$. We will refer back to these later.

In the relativistic $(k_B T >> mc^2)$ and $k_B T >> \mu$ limit for bosons

$$\begin{aligned} \epsilon &= \frac{\pi^2}{30\hbar^3 c^3} g(k_B T)^4 \\ n &= \frac{\zeta(3)}{\pi^2 \hbar^3 c^3} g(k_B T)^3 \\ P &= \epsilon/3. \end{aligned} \tag{1.19.13}$$

where ζ is the Riemann Zeta function and $\zeta(3) = 1.202....$

In the relativistic and $k_B T >> \mu$ limit for fermions we get

$$\begin{aligned} \epsilon &= \frac{7}{8} \frac{\pi^2}{30\hbar^3 c^3} g(k_B T)^4 \\ n &= \frac{3}{4} \frac{\zeta(3)}{\pi^2 \hbar^3 c^3} g(k_B T)^3 \\ P &= \epsilon/3. \end{aligned} \tag{1.19.14}$$

In the non-relativistic and dilute ($f << 1$ for most particles) limits we can neglect the ± 1 factor in the denominator of f so we get the same result for both bosons and fermions:

$$\begin{aligned} n &= g \left(\frac{mc^2 k_B T}{2\pi\hbar^2 c^2} \right)^{3/2} \exp\left[-(mc^2 - \mu)/(k_B T)\right] \\ \epsilon &= mc^2 n \\ P &= n k_B T << \epsilon. \end{aligned} \tag{1.19.15}$$

Homework

19.1: Starting from the appropriate integrals of the phase space distribution function over momentum space show that for relativistic massless bosons with $\mu = 0$ that $P = \epsilon/3$. [Note: this is a result you've seen before.]

19.2: The gas in this room consists of non-relativistic particles so to a very good approximation $E(p) = mc^2 + p^2/(2m)$. Derive this approximation from $E^2 = p^2 c^2 + m^2 c^4$.

19.3: Derive the Maxwell-Boltzmann distribution of velocities for the gas in the room (assuming for simplicity it's a gas of one type of particle (which of course it is not)); i.e., derive this: the probability that a particular gas particle has a speed between v and $v + dv$ is proportional to $v^2 \exp(\frac{-mv^2}{2k_B T}) dv$ where m is the mass of the gas particle.

Also for the gas in the room, for the huge majority of particles, the exponential term is much greater than one so you can ignore the ± 1 in the denominator and approximate $f = 1/[\exp(E-\mu)/kT]$.

19.4: Find the number density of particles as a function of μ, m and T assuming they are in kinetic equilibrium and that they are non-relativistic and you can neglect the ± 1 term in the denominator. It's OK to leave an unevaluated integral in your answer, but simplify it as much as possible and make a change of variables so the integration variable is dimensionless, and so the integral is just a number; i.e., it does not have any dependence on T, μ or m.

This page titled 1.19: Equilibrium Statistical Mechanics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

- **Current page** by Lloyd Knox is licensed CC BY 4.0.
- **2.16: S20 Equilibrium Statistical Mechanics SOLUTIONS** by Lloyd Knox is licensed CC BY 4.0.

1.19.1: Chapter 19 footnotes

1. There is *not* agreement between predictions and observations of Lithium abundance, something known as the "Lithium problem."

2. This may look strange, but it's generally true. Remember that in the non-relativistic limit $E = mc^2 + p^2/2m \simeq mc^2$.

This page titled 1.19.1: Chapter 19 footnotes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.20: Equilibrium Particle Abundances

At sufficiently high temperatures and densities, reactions that create and destroy particles can become sufficiently rapid that an equilibrium abundance is achieved. In this chapter we assume that such reaction rates are sufficiently high and work out the resulting abundances as a function of the key controlling parameter $\frac{mc^2}{k_BT}$. We will thus see how equilibrium abundance changes as the universe expands and cools. We will do so for the specific case of a fermionic particle (χ) and its anti-particle ($\bar{\chi}$) with non-zero mass (rest mass) m, and $g=2$, but the generalization to zero rest mass, bosons, and/or arbitrary g is trivial.

We make a few additional assumptions:

1. $n_\chi = n_{\bar{\chi}}$ initially (perhaps because these number densities are zero initially),
2. Any production or destruction of χ includes a production or destruction (respectively) of $\bar{\chi}$.
3. The reactions $\chi + \bar{\chi} \leftrightharpoons 2\gamma$ are fast.
4. Reactions that create and destroy photons are fast, such as $e^- + p^+ \rightarrow e^- + p^+ + \gamma$.

Assumption (4) allows us to determine that the photons have zero chemical potential; i.e., $\mu_\gamma = 0$.

Assumption (3) provides us with a constraint on μ_χ and $\mu_{\bar{\chi}}$: $\mu_\chi + \mu_{\bar{\chi}} = 2\mu_\gamma = 0$.

Assumption (2) ensures that the initial equality in Assumption (1) persists over time, providing us with another constraint on the chemical potentials. We can find this constraint with the following argument. The number density of a particle in kinetic equilibrium is determined entirely by its number of internal degrees of freedom, g, its mass, the temperature, and the chemical potential. The particle and antiparticle are able to exchange kinetic energy (with themselves, as well as other particles) and so share the same temperature. They also have the same mass and number of internal degrees of freedom. Therefore, the only thing left that affects the number density that they conceivably do not have in common is their chemical potentials. Since their number densities are equal, their chemical potentials must be equal.

So we simultaneously have $\mu_\chi = \mu_{\bar{\chi}}$ and $\mu_\chi + \mu_{\bar{\chi}} = 0$. The only solution is $\mu_\chi = \mu_{\bar{\chi}} = 0$.

We now have all the parameters of the phase space distribution function pinned down except for the temperature, T, so we are now ready to calculate the number density as a function of T. We have for fermions with zero chemical potential:

$$n_\chi = n_{\bar{\chi}} = \frac{g}{h^3}\int d^3p \left[\exp\left(\frac{E(p)}{k_BT}\right)+1\right]^{-1} \tag{1.20.1}$$

which we will now examine in the relativistic and then non-relativistic limits.

Relativistic Limit

Assuming $E(p) = pc$ we find

$$n_\chi = \frac{4\pi g}{h^3}\left(\frac{k_BT}{c}\right)^3 \int_0^\infty x^2 dx \left[e^x+1\right]^{-1}. \tag{1.20.2}$$

The integral can be numerically evaluated, or looked up in an integral table, with the result that it is $\frac{3}{2}\times\zeta(3) \simeq \frac{3}{2}\times 1.202$, where ζ is the Riemann zeta function and $\zeta(3) = \Sigma_1^\infty \frac{1}{n^3}$. We thus find

$$n_\chi = 6\pi\zeta(3) g\left(\frac{k_BT}{hc}\right)^3. \tag{1.20.3}$$

Box 1.20.1

Exercise 20.1.1: Derive

$$n_\chi = \frac{4\pi g}{h^3}\left(\frac{k_BT}{c}\right)^3 \int_0^\infty x^2 dx \left[e^x+1\right]^{-1} \tag{1.20.4}$$

from

$$n_\chi = \frac{g}{h^3}\int d^3p \left[\exp\left(\frac{E(p)}{k_BT}\right)+1\right]^{-1} \tag{1.20.5}$$

and $E(p) = pc$. Use the transformation to spherical momentum coordinates to rewrite $\int d^3p$ as $4\pi \int dpp^2$ and then transform the integration variable via $x = pc/(k_B T)$.

It is often helpful to look at the comoving number density $\equiv a^3 n$ because this quantity will be fixed as expansion occurs unless there is net creation or destruction of particles. Examining the comoving number density allows us to highlight changes that are due to effects other than simple dilution due to increased volume. Assuming $T \propto 1/a$ we find the comoving number density is *independent of temperature*, since $T^3 a^3$ is independent of temperature. Even though particles are rapidly being created and destroyed, the net result is that the number density has the same dependence on the scale factor, $n_\chi \propto a^{-3}$ that would be the case if there were no creation or destruction.

Non-relativistic Limit

In the non-relativisitic limit the kinetic contribution to the square of the energy p^2c^2 is much less than the rest-mass contribution to the square of the energy m^2c^4. So we have

$$\begin{aligned} E &= \sqrt{m^2c^4 + p^2c^2} = mc^2\sqrt{1 + p^2/(m^2c^2)} \\ &\simeq mc^2(1 + p^2/(2m^2c^2)) = mc^2 + p^2/(2m) \end{aligned} \tag{1.20.6}$$

and therefore

$$n_\chi = \frac{4\pi g}{h^3}\int_0^\infty p^2 dp \left[\exp[\frac{mc^2 + p^2/(2m)}{k_B T}] + 1\right]^{-1}. \tag{1.20.7}$$

In the non-relativistic regime $mc^2 >> k_B T$ (since in the non-relativistic regime $k_B T$ is a typical particle kinetic energy), so we can neglect the +1 in the phase-space distribution function, which allows us to pull $\exp[mc^2/(k_B T)]$ out of the integral, and make the variable substitution $x = p/\sqrt{2mk_B T}$ so that

$$n_\chi = \frac{4\pi g}{h^3}\exp\left(\frac{-mc^2}{k_B T}\right)(2mk_B T)^{3/2}\int_0^\infty x^2 dx e^{-x^2}. \tag{1.20.8}$$

The integral is $\sqrt{\pi}/4$. Using that and with some rearranging we get

$$n_\chi = \frac{(2\pi)^{3/2} g}{h^3 c^3}(k_B T)^3 \exp\left(\frac{-mc^2}{k_B T}\right)\left(\frac{mc^2}{k_B T}\right)^{3/2}. \tag{1.20.9}$$

Evaluation of the comoving number density brings in a factor of a^3 that cancels out the first T-dependent factor and we get:

$$a^3 n_\chi \propto \exp\left(\frac{-mc^2}{k_B T}\right)\left(\frac{mc^2}{k_B T}\right)^{3/2}. \tag{1.20.10}$$

We thus see that the abundance of the particles and antiparticles (recall $n_\chi = n_{\bar{\chi}}$) is controlled by $mc^2/(k_B T)$ and is exponentially suppressed when this quantity is much greater than 1.

Box 1.20.2

Exercise 20.2.1: Make a log-log sketch of $mc^2/(k_B T)$ vs. $a^3 n_\chi$ assuming the transition between the two regimes (relativistic and non-relativistic) is as smooth as possible. Explicitly identify $mc^2/(k_B T) = 1$ on the x-axis and the relativistic and non-relativistic regimes. Indicate which direction along the x-axis (left or right) corresponds to increasing time and scale factor.

This page titled 1.20: Equilibrium Particle Abundances is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.21: Hot and Cold Relics of the Big Bang

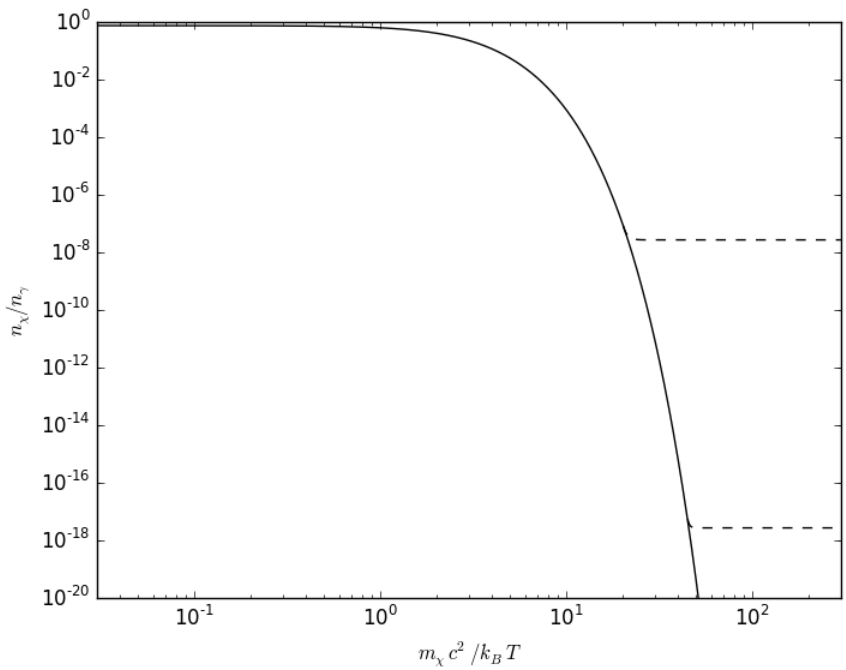
In the previous chapter we worked out the abundance of a massive fermionic particles species, χ and antiparticle $\bar{\chi}$ under a particular set of assumptions including kinetic equilibrium as well as chemical equilibrium maintained by reactions that set the chemical potentials $\mu_\chi = \mu_{\bar{\chi}} = 0$. We saw that, given these assumptions, if we artificially set the initial values of n_χ and $n_{\bar{\chi}}$ to zero we would, at early times when $m_\chi c^2 \ll k_B T$, rapidly evolve to an abundance of them very similar to that of the photons -- only differing by the number of degrees of freedom, g, and the slight difference that arises due to the small difference in the form of f for fermionic and bosonic species. As long as all our assumptions remain valid, the abundance of the species would eventually start to decrease as $k_B T$ dropped below $m_\chi c^2$, suppressed by the Boltzmann factor $\exp(-m_\chi c^2/(k_B T))$. In this scenario, assuming that today $m_\chi c^2 \gg k_B T$, we would have negligible amounts of these particles still around today.

Particles that at one time were driven to kinetic and chemical equilibrium abundances in the big bang, and survive to today are called "thermal relics." Survival to today sometimes require a departure from the equilibrium abundances worked out in the previous chapter. These departures, in some cases, occur because the reactions that maintained chemical equilibrium become too slow to continue to do so as the temperature drops and the equilibrium abundance changes. We refer to the process of the reactions becoming slow as "freeze-out." If freeze-out happens when the particles are still relativistic we call them "hot relics," if it occurs when they are non-relativistic we call them "cold relics." In this chapter we discuss hot and cold relics of the big bang.

Freeze-out

Calculation of reaction rates, and a complete treatment of non-equilibrium abundance evolution, is beyond the scope of this course. Here we summarize some important results:

- When per-particle reaction rates, Γ, are much greater than the expansion rate, H, then the abundance is rapidly driven to an equilibrium amount with chemical potentials governed by the reactions in question.
- The ratio Γ/H in general decreases over time as the universe expands and therefore temperature and density drops. Thus reactions tend to go from being "fast" (able to maintain equilibrium) to being "slow" (not able to maintain equilibrium). Note that both H and Γ drop with time, but Γ usually drops more rapidly.
- When $\Gamma << H$ for reactions that create and destroy χ and $\bar{\chi}$ particles then it is a good approximation to ignore these processes and assume their numbers just dilute with the expansion. We therefore have $n_\chi \propto a^{-3}$ and $n_{\bar{\chi}} \propto a^{-3}$.
- In fact, a fairly good approximation is to define a freeze-out temperature T_F to be the temperature at which $\Gamma = H$ and then assume the equilibrium abundance for $T > T_\mathrm{F}$ and $n_\chi \propto a^{-3}$ for $T < T_\mathrm{F}$, with the proportionality constant chosen so as to have a continuous abundance curve at $T = T_\mathrm{F}$.



Solid curve: Equilibrium abundance for a massive fermionic species with $g = 2$ and zero chemical potential relative to massless bosonic species with $g = 2$. This is very similar to what you were asked to sketch in the previous chapter since $n_\gamma \propto a^{-3}$. Dashed curves: the same ratio for two cold relics over time, with differing values of T_{F}. These are considered cold relics since they freeze out at $m_\chi c^2 > k_B T$. Note that the later the occurrence of freeze-out, the lower the relic abundance. Not shown: hot relics, which would freeze out at $T < T_{\mathrm{F}}$. Note that for hot relics the relic abundance is independent of when freeze-out occurs since the equilibrium curve is flat there.

Hot Relics: Photons and Neutrinos

At $z > 10^7$ or so, corresponding to temperatures greater than about $k_B T = 4$ keV reactions such as

$$e^- + p^+ \longleftrightarrow e^- + p^+ + \gamma \tag{1.21.1}$$

that create or destroy photons become slow. Since this happens while photons are relativistic (since photons are always relativistic), photons qualify as hot relics. This is the relic background predicted by Alpher and Herman in the 1940s and serendipitously discovered by Penzias and Wilson in 1964. We know it today as the cosmic microwave background (CMB). With a temperature of about 2.73°K the intensity of this light peaks in the microwave region of the spectrum. The spectrum has been measured with high precision and found to be consistent with a black body.

Various different types of distortion away from black body have been constrained by these measurements, including a non-zero chemical potential. Using data from the FIRAS instrument on the COBE satellite, cosmologists have placed a limit of $\mu/(k_B T) < 10^{-5}$. This level of consistency with a black body spectrum places constraints on possible scenarios in which energy is injected into the plasma of the big bang, for example from some particle species that decays while out of equilibrium into two photons. Such injection of energy would heat up the plasma. If the energy injection occurred at $z > 10^7$, the photon-number-changing reactions would ensure that the chemical potential remained zero, with the result that we would see no evidence in the spectrum of the CMB today. If the energy injection occurred at lower redshifts then the lack of photon-number-changing reactions would lead to a non-zero chemical potential. With a sufficient amount of energy injected, this distortion of the spectrum away from that of a black body would be discernible with current data.

Neutrinos are subatomic particles first inferred from examination of radioactive decay products. For example, a free neutron will decay to an electron, a proton and a neutrino. If one measures the momentum of the electron and proton from a neutron decaying at rest, one finds that either energy and momentum conservation is violated, or there must be an unseen additional decay product: the neutrino. There are three types of neutrinos in the standard model of particle physics, one paired with each of the three charged

leptons in the standard model. Together these are the electron and the electron neutrino, the muon and the muon neutrino, and the tauon and the tau neutrino.

Unlike the charged leptons, which interact with other particles via both the weak and electromagnetic forces, neutrinos only interact via the weak force. This force is appropriately named: the weak interactions are very weak. Neutrinos produced in nuclear reactions in the sun stream straight out of the sun hardly interacting at all, whereas photons produced in nuclear reactions scatter around the plasma of the sun for millions of years before finally making it to the surface. Despite the weakness of the interactions, at sufficiently early times the universe was hot and dense enough that reactions that produce and destroy neutrinos were occurring rapidly, as well as neutrino scattering reactions that exchange kinetic energy. The universe was in this state at temperatures $k_B T > 0.8$ MeV. Neutrinos were highly relativistic at this time, and thus they qualify as hot relics.

We have strong indirect indications of the existence of a relic background of cosmic neutrinos. Due to the weakness of their interactions, their low energy today (having been cooled by the expansion), and the drop in reaction rates with energy, it is exceedingly difficult, and perhaps practically impossible, to directly detect the cosmic neutrino background.

When we study Big Bang Nucleosynthesis (BBN) we will see some of the indirect evidence for the neutrino background. The energy density of the cosmic neutrino background, ϵ_ν, contributes to the total energy density and thus, via the Friedmann equation, to the expansion rate. The existence of the neutrinos means that, at a given temperature, the expansion rate is faster than it would be otherwise. We will see that this increased expansion rate affects the abundances of light elements. We will study the case of helium-4 in particular.

I have a particular interest in neutrinos and the cosmic neutrino background. The neutrinos have a somewhat-unique influence on the oscillations of standing waves in the plasma of the big bang, which follows from the fact that they can stream through the plasma at the speed of light, unlike the photons that readily scatter off of free electrons. A result of this free-streaming is that gravitational potentials that drive the standing wave oscillations decay more rapidly than they would otherwise, fast enough that they alter the temporal phases of the oscillations, an effect that is observable in the statistical properties of the cosmic microwave background. My graduate students and I were the first to isolate this effect in the data. The paper is in Physical Review Letters. There is a popular account in Scientific American, and in some blog posts including this one, and this one which is somewhat overblown about the significance, but provides, as background material, a nice summary of our historical progress on understanding the big bang. The work also led to an article about me and my research and teaching in the Sacramento Bee.

Box 1.21.1

Exercise 21.1.1: At $z \simeq 10^7$ some reactions involving photons became slow, and at $z \simeq 10^3$ some other reactions involving photons became slow. Which epoch corresponds to freeze-out for photons? What is the significance of the later epoch?

Answer

Photon freeze-out is when reactions that change photon number become slow. This is at $z \sim 10^7$. The later epoch of $z \sim 10^3$ is when photons stop, for the most part, interacting with matter. This is when the universe transitions from an opaque plasma to a transparent neutral gas.

Exercise 21.1.2: A weaker interaction (slower production and annihilation rates) usually means freeze-out happens earlier or later?

Answer

The weaker the interaction, the earlier the reaction rates will become slow for the reactions that change the particle number, so the earlier freeze-out will occur.

Exercise 21.1.3: For hot relics, why is the abundance today relatively insensitive to freeze-out temperature?

Answer

Because for hot relics, $n \propto a^{-3}$ applies whether in equilibrium or out. In contrast, for cold relics, staying in equilibrium means the abundance faces Boltzmann suppression.

Exercise 21.1.4: Why do additional species of light particles potentially lead to a faster expansion rate in the early universe at a given temperature?

Answer

Because, at a given temperature, more degrees of freedom mean a greater energy/mass density and therefore greater expansion rate via the Friedmann equation $H^2 = 8\pi G\rho/3$.

Cold Relics: The WIMP Miracle (or Misleading Coincidence?)

We've learned so far that photons and neutrinos are produced thermally in the big bang. Potentially there are also particles produced in the big bang that are not in the standard model of particle physics. The cold dark matter that seems to dominate the mass density of the universe might be such a particle. Here we consider a general class of models of dark matter in which it is a Weakly Interacting Massive Particle (WIMP).

WIMPs have received a lot of attention. There are major experimental efforts underway to detect WIMPs indirectly (by seeing evidence of their annihilation into standard model decay products in regions of dense dark matter such as the galactic center or centers of dwarf galaxies) to detect them directly (via rare interactions of dark matter with baryonic matter in a sensitive detector), and to produce them in particle colliders such as the Large Hadron Collider in Europe.

This attention follows, at least in part, from a theoretical result called the "WIMP Miracle." The miracle is as follows. An idea in particle physics called "supersymmetry," which is attractive to particle physics because of problems it solves having nothing to do with cosmological observations, leads to a collection of new particles, one of which could quite naturally turn out to be the dark matter. If one works out reaction rates for these supersymmetric particles, and when freeze-out occurs for the lightest one (which is the only stable one), the lightest one can easily have (depending on choices of free parameters) the right relic abundance today. More generally, supersymmetry offers a solution to the "hierarchy" problem of particle physics, a problem associated with the weak interaction scale. Weak interaction particle cross sections are about what one needs in order to end up with relic abundances today roughly consistent with the mass density of dark matter as inferred from cosmological observations.

What is the relationship between particle cross sections and their relic abundance? Particle interaction cross sections control reaction rates. The smaller the cross sections, the slower the reaction rates. The slower the reaction rates, the earlier freeze-out occurs. The earlier freeze-out occurs (not in terms of time itself, but in terms of the time-like variable $m_\chi c^2/(k_B T)$), the higher the resulting abundance. A weak interaction level of cross section (which the supersymmetric particles generally have) turns out to be just right for getting the right density of dark matter today. This is the WIMP miracle.

Or... it could be just a coincidence that has thrown us off the right path toward finding out what the dark matter really is. So far, searches for supersymmmetric dark matter, after decades of searching and improving sensitivity, have resulted in upper limits. There are some claimed detections, but none of these have yet been reproduced by other experiments.

Box 1.21.2

Exercise 21.21: For cold relics, why are weaker interactions associated with higher relic abundances?

Answer

Weaker interactions means earlier freeze out when there has been less Boltzmann suppression.

Exercise 21.2.2: If there is a new particle physics theory that has in it new particles, or new interactions, or both, why is it important to consider cosmological consequences? What might they be?

Answer

The particles might be thermally produced in the big bang. If they are stable, they will contribute to the mass density, and therefore the expansion rate, over time. They could potentially change the predictions of big bang nucleosynthesis.

Homework

21.1: Starting when $k_BT \simeq 0.511$ MeV, (so that $m_ec^2/(k_BT) = 1$) the number density of electrons and positrons, relative to photons, begins to drop. Eventually, almost all the positrons and electrons are gone. This whole process occurs after the neutrinos have decoupled (at $k_BT = 0.8$ MeV) so the annihilations of the electrons and positrons all goes into heating up the photons (some goes into the remaining electrons and nuclei, but their number densities and kinetic energy densities are tiny compared to photon number an energy densities). Assuming that the process does not change the entropy in a comoving region, show that the end result is that neutrinos are cooler relative to photons by an amount

$$T_\nu/T_\gamma = \left(\frac{4}{11}\right)^{1/3}. \tag{1.21.2}$$

You'll need to use the fact that the entropy density (physical density, not comoving density) for a relativistic species is

$$s = \frac{2\pi^2}{45} g k_B \left(\frac{k_BT}{\hbar c}\right)^3 \tag{1.21.3}$$

for bosons and

$$s = \frac{7}{8}\frac{2\pi^2}{45} g k_B \left(\frac{k_BT}{\hbar c}\right)^3 \tag{1.21.4}$$

for fermions.

Some hints:

1) You have a conserved quantity. Evaluate it at a time when electrons and positrons are relativistic and at a later time when the electrons and positrons have all disappeared. Use this to figure out how the early temperature is related to the later temperature.

2) Since the neutrinos stop interacting with themselves and with electrons, positrons, and photons, their entropy is conserved separately from the electrons, positrons, and photons. The entropy of the electron, positron, and photon system is also conserved.

3) Electrons are spin 1/2 so have $g = 2$. Same for positrons.

4) Initially neutrinos, electrons, positrons, and photons are exchanging kinetic energy and so their initial temperatures (prior to the electrons and positrons starting to disappear) are equal.

21.2: From atmospheric neutrino oscillations, we know that the lightest that the most massive neutrino can be is 0.048 eV/c^2. Assuming it's as light as possible, what is a typical speed for the most massive neutrinos in the cosmic neutrino background today? (I'm looking for something correct to within a factor of 2 or so). Keep in mind that neutrinos freeze out while still relativistic so they keep $f = [\exp(pc/k_BT) + 1]^{-1}$ even though that is not equal to $[\exp(E/(k_BT)) + 1]^{-1}$. For such a distribution a typical value of pc is k_BT.

21.3: The number density of cosmic microwave background photons today is about 400/cm^{-3}. Assuming there are 3 species of neutrino, each with $g = 2$, and that their temperature is reduced relative to photons as described in problem 22.1, what is the number density of cosmic neutrinos today?

21.4: Combining results from 21.2 and 21.3, give an estimate of the number of highest-mass cosmic neutrinos flowing through you per second.

21.5: The baryon-to-photon ratio n_b/n_γ is about 6×10^{-10}. Assuming, for simplicity, that all the baryons are protons, calculate the temperature at which there is one photon with energy above 13.6 eV for every baryon. You may use this result (which you could obtain by integrating over the photon energy distribution): the fraction of photons with energy above E is

$$\frac{n(>E)}{n} \simeq \left(\frac{E}{k_BT}\right)^2 \exp\left(\frac{-E}{k_BT}\right) \tag{1.21.5}$$

for $E >> k_BT$.

This page titled 1.21: Hot and Cold Relics of the Big Bang is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

- **Current page** by Lloyd Knox is licensed CC BY 4.0.
- **2.18: S22. Hot and Cold Relics of the Big Bang** by Lloyd Knox is licensed CC BY 4.0.

SECTION OVERVIEW

1.22: Overview of Thermal History

** under construction **

We should have a graphic here with time on the top axis and scale factor on the bottom axis, and mass density on the y axis. Label epochs such as double Compton freeze-out, e-/e+ annihilation, quark-hadron phase transition, weak interaction freeze-out, BBN, matter-radiation equality, and recombination. Maybe we stop a little ways after recombination.

We could have three such graphics, one with speculative very early-universe stuff (starting with inflation and reheating), one from quark-hadron phase transition to recombination, and one from first stars and quasars to present day.

Topic hierarchy

This page titled 1.22: Overview of Thermal History is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.23: Big Bang Nucleosynthesis - Predictions

Overview

Big Bang Nucleosynthesis (BBN) is the process by which light elements formed during the Big Bang. The agreement between predicted abundances and inferences from observations of primordial (pre-stellar) abundances is a major pillar of the theory of the hot big bang and reason we can speak with some confidence about events in the primordial plasma in the first few minutes of the expansion. Elements created at these very early times include Deuterium, Helium-3, Lithium-7, and, most abundantly, Helium-4. In this chapter we focus on the theory of Helium-4 production.

State of the art calculation of Helium-4 production in the Big Bang involves following a fairly large reaction network, between the various light elements in their thermal bath of photons and neutrinos. It is a sufficiently complicated calculation that it is done numerically on computers. Here we present analytic arguments that help us to understand why the results come out the way they do. This is a common situation in physics -- although perhaps not so common in physics as it is taught to undergraduates. Most problems are way too hard to solve analytically from first principles. Some problems are amenable to being solved numerically. When problems are solved numerically, we often want more than to just know the result of the calculation. We want to understand *why* the result is what it is. Such understanding is valuable for our own satisfaction, but also it often allows us to figure out, at least qualitatively, what the result will be if some assumption is changed. It is useful to be able to do this, rather than have to re-do the numerical calculation every time you get curious about the result of changing some input to the calculation.

In this chapter we present the results of numerical calculations of light element production in the big bang. We also provide analytic insight into why Helium production works out the way it does. We then use that analytic insight to understand how He-4 abundances are sensitive to conditions in the Big Bang.

With standard assumptions about the Big Bang, we find that about 25% of the mass that is in baryons ends up in Helium-4. In this chapter we will examine, as the basis of our analytic understanding of this result, the following sequence of events:

- At $k_BT \geq 0.8\mathrm{MeV}$, neutrons and protons are in chemical equilibrium.
- At $k_BT \simeq 0.8\mathrm{MeV}$ their ratio freezes-out at $\frac{n_n}{n_p} \simeq \frac{1}{5}$.
- ^{4}He production proceeds through the intermediate step of Deuterium (D) production, which is inhibited until $k_BT \simeq 0.06\mathrm{MeV}$.
- At this temperature, nearly all neutrons end up in ^{4}He, but by this point neutron decay has reduced $\frac{n_n}{n_p}$ from $\frac{1}{5}$ to $\frac{1}{7}$.

After presenting the story of Helium-4 production in the standard cosmological model, we discuss the sensitivity of the Helium-4 abundance to assumptions such as the value of Newton's constant, G.

Conditions prior to neutron-proton freeze-out

We begin by considering the universe at a time just prior to the BBN epoch, when the temperature of the universe is such that $k_BT \simeq$ 10 MeV, much hotter, and denser, than it is today. At this time, the Universe is dominated by radiation, including paired relativistic particles like e^- and e^+ and ν and $\bar{\nu}$.

However, at this same time, baryon kinetic energies have dropped enough that they are nonrelativistic ($k_BT \ll m_pc^2$). The universe is still hot and dense enough to keep weak interactions such as these:

$$n + \nu_e \longleftrightarrow p + e^- \tag{1.23.1}$$

$$n + e^+ \longleftrightarrow p + \bar{\nu}_e \tag{1.23.2}$$

rapid. These reactions keep the neutrons and protons in chemical equilibrium with:

$$\mu_n + \mu_{\nu_e} = \mu_p + \mu_{e^-} \tag{1.23.3}$$

$$\mu_n + \mu_{e^+} = \mu_p + \mu_{\bar{\nu}_e} \tag{1.23.4}$$

In kinetic equilibrium, electrons and positrons are at the same temperature, and they have the same masses. Also $n_{e^-} \simeq n_{e^+}$, which implies

$$\mu_{e^-} \simeq \mu_{e^+} \tag{1.23.5}$$

$e^- + e^+ \longleftrightarrow 2\gamma$ reactions are fast, so

$$\mu_{e^+} + \mu_{e^-} = 0 \tag{1.23.6}$$

Using 1.23.5 and 1.23.6, we can therefore conclude that their chemical potentials are near zero. Similar arguments apply to neutrinos, although, we don't have good evidence that $n_\nu = n_{\bar{\nu}}$. However, in the standard cosmological model, we assume $n_\nu = n_{\bar{\nu}}$, and hence conclude that $\mu_\nu = \mu_{\bar{\nu}} = 0$.

Box 1.23.1

Exercise 23.1.1: Based on the assumptions presented so far show that $\mu_n = \mu_p$.

Neutron-Proton Freeze-out

When the temperature drops to T_f, the critical temperature at which weak interactions are no longer fast, the neutron to proton ratio "freezes-out" to a nearly constant value. From the number density equations shown at the end of chapter 20 and from our previous result $\mu_n = \mu_p$, we can find the equilibrium ratio of neutrons to protons to be:

$$\frac{n_n}{n_p} = \left(\frac{m_n}{m_p}\right)^{3/2} \exp\left[-\frac{(m_n - m_p)c^2}{k_B T_f}\right] \tag{1.23.7}$$

Box 1.23.2

Exercise 23.2.1: You are given number density equations in chapter 20. Use them to derive Eq. 1.23.7. Assume that reactions 1.23.1 and 1.23.2 are fast.

After plugging in the known quantities and the correct freeze-out temperature ($k_b T_f \approx 0.8\text{MeV}$), this fraction evaluates to:

$$\frac{n_n}{n_p} \simeq \exp\left(-\frac{1.3\text{MeV}}{0.8\text{MeV}}\right) \approx \frac{1}{5} \tag{1.23.8}$$

where we have used the fact that $\frac{m_n}{m_p} \simeq 1$ and $(m_n - m_p)c^2 = 1.3$ MeV. This ratio slowly drops over time due to neutron decay. Neutrons decay as $n \longrightarrow p + e^- + \bar{\nu}$ with $t_{1/2} = 610s$.

Nuclear Statistical Equilibrium

By the time the universe has cooled to $kT \simeq 0.3$ MeV there is a sense in which all nuclear matter "wants" to be the bound state of 2 neutrons and 2 protons, helium-4, or ^{4}He. But because the reactions necessary to create ^{4}He are too slow, this does not happen for a while yet.

To explain what we mean by this "wanting" we introduce the concept of nuclear statistical equilibrium (NSE). NSE applies when all the reactions that can change the numbers of the different nuclei are sufficiently rapid. Here are some of these reactions, starting with one that forms deuterium, the bound state of a neutron and proton

$$p + n \longrightarrow D + \gamma \tag{1.23.9}$$

and then those that produce ^{3}He

$$D + p \longrightarrow \ ^3He + \gamma \tag{1.23.10}$$

$$D + D \longrightarrow \ ^3He + n \tag{1.23.11}$$

those that produce ^{3}H

$$D + D \longrightarrow \ ^3H + p \tag{1.23.12}$$

$$^3He + n \longrightarrow \ ^3H + p \tag{1.23.13}$$

and those that produce ^4He

$$D + D \longrightarrow \ ^4He + \gamma \tag{1.23.14}$$

$$^3He + D \longrightarrow \quad ^4He + p \tag{1.23.15}$$

$$^3H + D \longrightarrow \quad ^4He + n. \tag{1.23.16}$$

We could continue listing these reactions right through to the production of the very heaviest elements, but we'll stop here. If they, and their associated reverse reactions, were all happening sufficiently rapidly, then that creates a set of relationships between their chemical potentials. The result is that, just like we see with Eq. 1.23.7 their abundance ratios would only depend on their masses, the temperature, and the baryon-to-photon ratio. For a value of the baryon-to-photon ratio that appears to be consistent with our universe, the result is that, if NSE were to obtain, just about all the mass in baryons would be in 4-He once the temperature falls below $kT \simeq$ 0.3 MeV. As the universe cools further, the most abundant element switches from Helium to Carbon, and as it cools further it keeps switching to heavier and heavier elements. Well before we get to today's temperature, if NSE obtained the whole way, all the baryons in the universe would be in ^{56}Ni, the nucleus with the highest per-baryon binding energy.

But this never happens, because the necessary reactions become slow, as we saw earlier, at temperatures as high as $kT \simeq$ 0.8 MeV. The main challenge to actually producing Helium-4 comes from the slowness of reaction 1.23.14, which is slow due to the small abundance of deuterium, which in turn is slow due to what we call "the deuterium bottleneck."

The Deuterium Bottleneck

The binding energy of a proton and neutron is $\sim 2.2\mathrm{MeV}$ and thus we would expect that at the time when the universe has cooled to $k_BT \simeq 2.2\mathrm{MeV}$ we would begin to see deuterium. However, we don't see deuterium in abundance until a much lower temperature of $k_BT \simeq 0.06\mathrm{MeV}$. This is due to the massive dominance of photon densities over baryon densities and the high energy tail in their distribution. Even if only a very small percentage of photons has enough energy to break deuterium bonds, that small percentage can still correspond to a greater number density than the total number density of deuterons. These number densities do not become equal until around $k_BT \simeq 0.06\mathrm{MeV}$. At this temperature enough Deuterium can form and survive for long enough to be converted into helium-4 by reaction 1.23.14. The deuterium bottleneck is broken.

Helium Production

Once the deuterium bottleneck is broken, Helium abundance can begin to move appreciably toward its very large NSE-desired abundance. The abundance of helium-4 increases until just about all of the available neutrons have been consumed. Without more neutrons, no more deuterium can be made, and there is once again no viable path for creating helium-4. Neutrons are thus the limiting fuel. Their abundance at $kT = 0.06$ MeV can be used to approximately determine the final abundance of helium-4. That abundance has decreased some since neutron-proton freezeout since the age of the universe at that temperature is about 340 seconds and the half life of the neutron is about 610 seconds.

$$\frac{n_n}{n_p} \simeq \frac{1}{5}\exp\left(-\frac{340s \times \ln 2}{610s}\right) \simeq \frac{1}{7} \tag{1.23.17}$$

Some of the products of the reactions we listed above go on to form ^7Li and others as well, but these account for only a very small fraction of the total baryonic mass.

Numerical evolution of the reaction network leads to the following predictions for primordial abundances relative to Hydrogen:

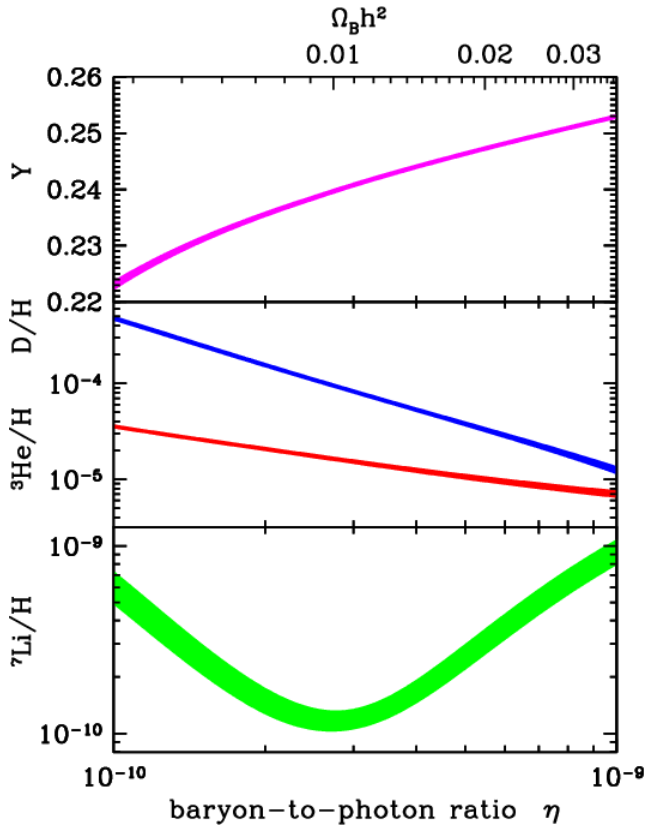


Figure 1.23.1: A Schramm Plot (Cyburt, Fields & Olive 2003) shows abundance predictions for standard BBN as a function of the baryon-to-photon ratio (bottom x axis) and $\Omega_b h^2$ (top x axis). The width of the curves indicates the uncertainty in the theoretical predictions, mostly due to uncertainties in nuclear reaction rates. All of the predictions are for abundances relative to hydrogen, except for helium which is expressed as Y, the fraction of baryonic mass in helium.

From the figure, we can ascertain some values for the primordial abundances of light elements. At a baryon-to-photon ratio value of $\eta = (6.14 \pm 0.25) \times 10^{-10}$, we have abundance predictions of about:

$$\frac{D}{H} \simeq 2.75 \times 10^{-5} \tag{1.23.18}$$

$$\frac{^3He}{H} \simeq 9.28 \times 10^{-6} \tag{1.23.19}$$

$$\frac{^7Li}{H} \simeq 3.82 \times 10^{-10} \tag{1.23.20}$$

These are all very small because essentially all neutrons go into creating ^4He at 0.06MeV. Keeping this in mind, and using our (neutron-decayed) ratio of neutrons to protons, we can find Y, the fraction of baryonic mass in ^4He:

$$Y = \frac{0.5 \times 4 m_n N_n}{m_n(N_n + N_p)} = \frac{2}{1 + N_p/N_n} \simeq 0.25 = 25\% \tag{1.23.21}$$

where m_n is the mass of a "nucleon" -- either a proton or neutron. The numerator here makes sense because the mass of helium-4 is about 4 m_n and we get half a helium-4 for every neutron. The bottom is clearly the total baryonic mass, where the values for N_n and N_p are to be understood as those just prior to formation of Helium-4.

Box 1.23.3

Exercise 23.3.1: Prove this percentage to yourself. From $\frac{N_n}{N_p} \simeq \frac{1}{7}$, draw 2 rows, each with 7 protons and 1 neutron, each. Draw a circle containing both neutrons and 2 protons. This is a ^4He nucleus. If $m_n \approx m_p$, what percentage of your total mass is in the ^4He nucleus?

Box 1.23.4

Exercise 23.4.1: Sketch a timeline with these two key BBN events in it: neutron-proton freeze-out and the end of the deuterium bottleneck. Label them and specify the temperature. Specify the ratio of neutrons to protons at both of these events. Why has the ratio decreased by the time of the later event?

Box 1.23.5

Exercise 23.5.1: If Newton's constant G were for some reason larger during BBN than it is today, the expansion rate at a given temperature would be higher (due to the Friedmann equation). Qualitatively, how would this impact predictions for Helium abudnance? Would Y, the fraction of baryonic mass in Helium, go up or down?

This page titled 1.23: Big Bang Nucleosynthesis - Predictions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.24: Big Bang Nucleosynthesis - Observations

The build-up of the chemical elements

We usually think of heavy elements as a product of nucleosynthesis (formation of nuclei) in stellar fusion and supernovae. Over billions of years of stellar processing and an overall increase in heavy element abundances, the chemical abundances of stars become richer with elements besides hydrogen, such as helium, oxygen, and iron. In a given star, we typically see a direct correlation between its oxygen abundance and iron abundance. If a star is oxygen-rich, it is likely a newer star made from dust clouds containing heavier elements, and thus likely has a high iron content as well. We can see this relationship in Figure 1, which plots the oxygen and iron content of many stars. This observed relationship supports our hypothesis that these elements were formed together over time with stellar processing.

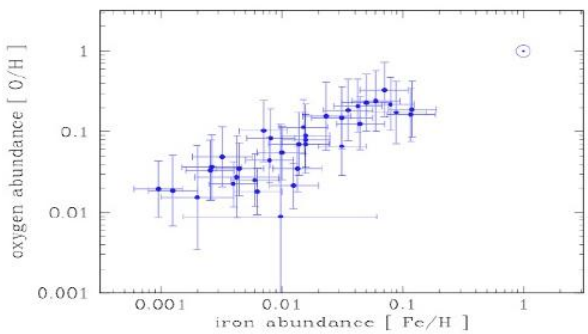
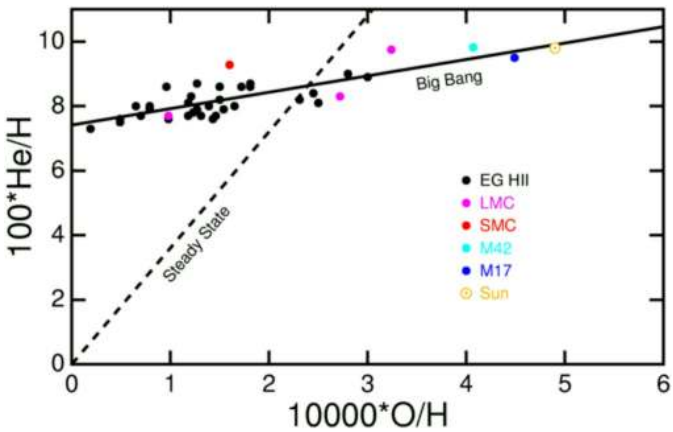


Figure 1.24.1: Relative abundances of oxygen-to-hydrogen and iron-to-hydrogen. We see a trend in which low-iron stars tend to have low-oxygen, while iron-rich stars are also oxygen-rich. This is because iron and oxygen were formed together over billions of years of stellar processing.

Observations of helium abundances gives us a different relationship, which can be seen in Figure 2. As with iron and oxygen, an oxygen-rich star is likely to contain more helium, which indicates that both helium and oxygen have been created over time with stellar processing. The difference is that we see a significant abundance of helium even in very old stars formed by gas clouds containing little to no heavy elements. This points to a primordial abundance of helium that existed even before stellar processing. Where did this helium come from?

Figure 1.24.2: Relative abundances of oxygen-to-hydrogen and helium-to-hydrogen in stars. While there is still a trend in which stars with low O/H also have lower He/H, even the most oxygen-poor stars contain a significant abundance of helium. This is because all these objects formed with a primordial abundance of helium, whereas there is essentially no primordial abundance of oxygen. (credit: Ned Wright)



In 1948, Gamow explored the very early universe as a source of elements heavier than hydrogen. He extrapolated Einstein's theory of an expanding universe with certain assumptions of what the universe is made of, and concluded that the universe was infinitely dense at a finite time in the past. He theorized that this early universe could be a prodigious source of heavier elements. If the universe began at infinitely high density and temperatures and underwent rapid expansion and cooling, atomic nuclei would form within a window of just a few minutes. Before this window, high-energy radiation did not permit any nuclei to survive, and after this window, temperatures were too low for the nuclear collisions to overcome the Coulomb repulsion.

Gamow's important discovery was that in order to avoid overproduction of helium and other heavy elements, the ratio of nucleons to photons had to be small. Since the number of photons in black body radiation is proportional to temperature cubed, this means that the "Big Bang" had to be very, very hot. From these parameters, Gamow theorized that we should see a background of heat and light from this period of high temperature and photon density. This background was discovered later in 1964 and is known as the Cosmic Microwave Background.

The Ashes of the Big Bang

Gamow's (and soon after, Ralph Alpher's) realization that chemical elements could be made during the early Universe paved the way for what is now referred to as Big Bang nucleosynthesis, which is the end-product of putting neutrons and protons in a hot, expanding universe. The original motivation of their work was to explain the origin of the abundance of all chemical elements, including hydrogen, helium, and all heavy elements. However, a few physical processes were neglected in their calculation, and as a result, the calculations by Alpher and Gamow produced orders of magnitude too many elements heavier than lithium, but roughly the correct amount of helium. In Chapter 23: Big Bang Nucleosynthesis - Predictions, we derived the amount of helium produced during the Hot Big Bang under some simplifying assumptions. Detailed calculations that include a full nuclear reaction network and all of the relevant physical processes can be used to estimate the complete set of nuclides that are created during these first few moments.

Only the lightest elements of the periodic table were made during Big Bang nucleosynthesis, including hydrogen (H), helium (He), and a trace amount of lithium (Li). The relative amounts of these nuclides depend on a competition between the expansion rate of the Universe and the rates of a network of nuclear reactions. This competition is parametrized by the relative number density of baryons to photons, commonly referred to as the baryon-to-photon ratio. Numerical calculations of these primordial nuclides have allowed us to understand the sensitivity of each primordial nuclide to the physics of the early Universe. For example, if there were an additional particle outside of the Standard Model, this might change the expansion rate of the Universe or interact with baryons in a way that changes the production and destruction of the primordial nuclides. Turning this statement around, if we could measure the relative abundance of the primordial elements, and combine this with a model of Big Bang nucleosynthesis, then we can learn about particle physics and cosmology a few minutes after the Big Bang.

The challenge is to find places in the Universe that have not significantly altered the primordial nuclides from their initial abundances set a few minutes after the Big Bang. This usually means that we have to find environments that have very few of the elements made by stars (e.g. oxygen, iron). Historically, the goal is to measure the number density of one primordial nuclide relative to hydrogen (i.e. the protons that are left over following Big Bang nucleosynthesis). While many nuclides are made during Big Bang nucleosynthesis, only the most abundantly produced primordial elements can currently be measured, including: deuterium (D, which is a heavy isotope of hydrogen), helium-3 (^{3}He), helium-4 (^{4}He), and lithium-7 (^{7}Li). We will discuss the current techniques and measurements of each of these primordial elements in turn.

Deuterium: A way to weigh the Universe

Deuterium is a heavy stable isotope of hydrogen, and Big Bang Nucleosynthesis makes just ~ 25 deuterons for every one million protons (a deuteron is the nucleus of a deuterium atom). The nucleus of deuterium contains one neutron and one proton, and it has very similar energy levels to the hydrogen atom. However, as a result of the neutron in the nucleus, the energy levels of a deuterium atom are shifted relative to that of hydrogen. This presents us with a problem, because if we want to measure the relative number of deuterium and hydrogen atoms of a pristine environment, we need to detect the atomic transitions of both deuterium and hydrogen. To estimate this shift, we can consider the Bohr model of the atom. The energy levels based on this model are given by:

$$\frac{1}{\lambda} = R\left(\frac{1}{n_f^2} - \frac{1}{n_i^2}\right)$$

where λ is the wavelength of the transition between levels n_f and n_i and R is the Rydberg constant of that atom. Since deuterium has an extra neutron in the nucleus, this changes the centre of mass of the atom, so the Rydberg constant is slightly different for deuterium and hydrogen. The ratio of the deuterium and hydrogen wavelengths is just the ratio of their Rydberg constants, which is simply the ratio of the reduced masses:

$$\frac{\lambda_D}{\lambda_H} = \frac{R_H}{R_D} = \frac{1 + \frac{m_e}{m_D}}{1 + \frac{m_e}{m_H}} \tag{1.24.1}$$

where m_e, m_H and m_D are the rest masses of the electron, proton and deuteron. The isotope shift, expressed as a velocity shift, is $c(\lambda_D - \lambda_H)/\lambda_H \simeq -81.6 \text{ km s}^{-1}$. So, if we want to detect deuterium and hydrogen transitions, we require that the environment must not contain Doppler motions that exceed this value.

In 1976, it was realized that there might be clouds of gas in the high redshift Universe that have very low Doppler motions, such that absorption by deuterium and hydrogen transitions could be imprinted on the light of a bright, unrelated background light source. The idea is similar to looking at a lighthouse on a foggy night. We know that there is fog between us and the lighthouse because we don't see all of the light from the lighthouse. The brightest "lighthouses" in the early Universe are quasars, which are rapidly accreting supermassive black holes in the centers of galaxies. By obtaining a spectrum of a quasar, we can study the clouds

of gas that are between our telescopes and the distant quasars. Furthermore, because the Universe is expanding, the absorption lines of all these gas clouds are redshifted by different amounts, leading to a plentiful forest of absorption lines (between a wavelength of $3500 - 4800$ in the example shown in Figure E); the absorption lines in this figure are almost entirely due to hydrogen atoms that are along the line-of-sight to the quasar.

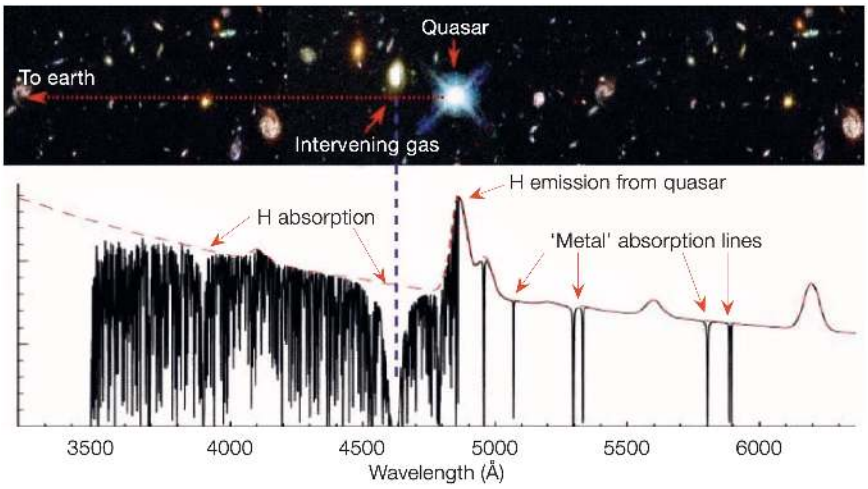


Figure $1.24.3$: A spectrum of a quasar is shown (black data). Galactic and intergalactic gas along the line-of-sight to this quasar absorb the quasar light at the wavelengths corresponding to atomic transitions. The expansion of the Universe causes these transitions to redshift relative to each other, leading to a forest of absorption lines, mostly comprised of hydrogen lines. There are also some absorption lines from heavy elements, called "metals", including oxygen and iron (credit: John Webb).

In some quasar spectra, there could be as many as several hundred hydrogen absorption lines, and all of these gas clouds have deuterium atoms in them as well. However, most of these gas clouds have Doppler motions that exceed the isotope shift of $81.6\ \mathrm{km\ s^{-1}}$, and cannot be used to measure the relative number of deuterium and hydrogen atoms. While this technique (called "quasar absorption line spectroscopy") was very promising, it took more than two decades of research and the advent of the 10 meter diameter Keck telescopes to find the first gas clouds where deuterium and hydrogen atomic transitions could be measured. An example of the deuterium and hydrogen absorption lines detected in a gas cloud using this technique is shown in Figure 4, where the transitions from from principal quantum number $n = 1$ to higher energy levels (a.k.a. the Lyman series of hydrogen and deuterium) are shown.

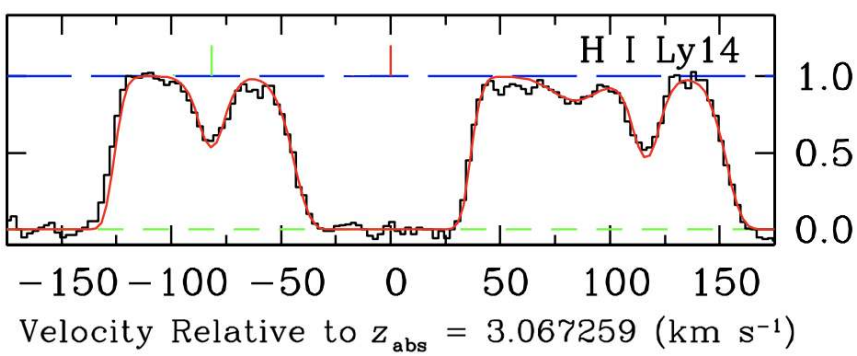
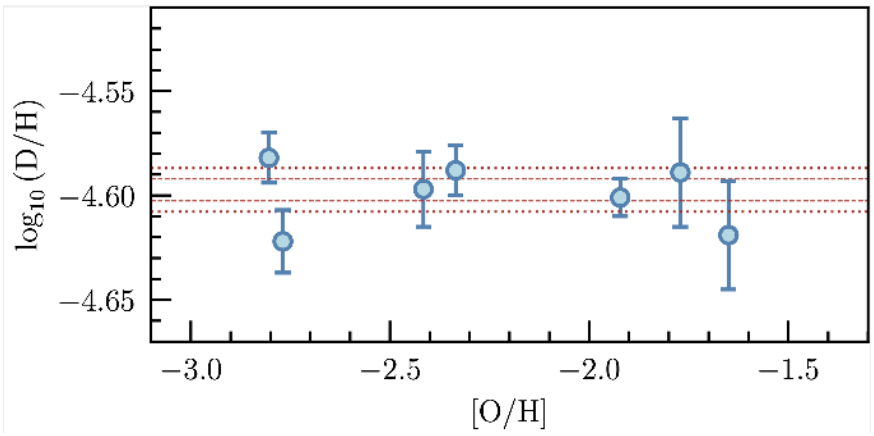


Figure $1.24.4$: The black data shows an example of a deuterium and hydrogen absorption line system. The deuterium absorption is indicated with the green line at $-81.6\ \mathrm{km\ s^{-1}}$, and the hydrogen line is shown with the red tick mark above the data. The red curve is a model fit to the data (black histogram). (credit: Ryan Cooke).

These same gas clouds also contain absorption lines from heavy elements (see Figure 3 for an example). This tells us that these gas clouds are not pristine reservoirs that have been untouched since the Big Bang; instead, their chemistry has been slightly altered by enrichment from stars. By measuring how many metals have been made relative to hydrogen in these environments, we can estimate how contaminated these environments are.

Since these first measurements were made, the technique used to identify these deuterium absorption systems has been refined. However, despite considerable work since the year 2000, the deuterium abundance has only been reliably and consistently measured for just seven gas clouds, demonstrating how challenging and rare this measurement is. These seven measurements are shown as the blue symbols in Figure 5, as a function of the amount of contamination by stars (represented by the oxygen abundance, [O/H]). Even though these seven independent gas clouds have had a different amount of stellar processing, they are all statistically consistent with each. This tells us that the gas clouds where these measurements were made have retained a primordial relative composition of deuterium and hydrogen, and their ratio is the same as the value set just minutes after the Big Bang.

Figure $1.24.5$: Seven independent measures of the deuterium abundance (i.e. the relative number of deuterium to hydrogen atoms) of gas clouds (blue symbols with error bars). The x-axis shows the oxygen abundance, displayed on a log scale relative to solar (i.e. -2 is one-hundredth of the number of oxygen atoms in the sun relative to hydrogen, while -3 represents one-thousandth of the



number of oxygen atoms). All seven measures are consistent with each other. The red horizontal lines represent 68 and 95 per cent confidence interval of the weighted mean value of these seven measures. (credit: Ryan Cooke).

Calculations of Big Bang nucleosynthesis indicate that the relative abundance of deuterium and hydrogen atoms is highly sensitive to the expansion rate of the Universe and the density of baryons (i.e. ordinary matter). If we assume the Standard Model of particle physics and cosmology, this sets the expansion rate of the Universe, and under this assumption, we can use the deuterium abundance to estimate the universal density of baryons. The baryon density is also a fundamental quantity that can be measured from the cosmic microwave background temperature fluctuations. Quite amazingly, the baryon density that is inferred from the seven measures of D/H agrees with the baryon density derived from the cosmic microwave background at one per cent precision. The astounding agreement of these two baryon density measures, based on completely independent physics, and based on two epochs of the Universe separated by almost 400,000 years, represents one of the strongest confirmations of our Standard cosmological model.

Helium-4 and the search for physics beyond the Standard Model

As discussed in Chapter 23: Big Bang Nucleosynthesis - Predictions, ^{4}He is the main nuclide produced during Big Bang nucleosynthesis. Calculations of Big Bang nucleosynthesis tell us that the amount of ^{4}He that is made depends primarily on the expansion rate of the Universe. There are three approaches to measure the primordial abundance of ^{4}He: (1) spectroscopic observations of metal-poor star-forming dwarf galaxies; (2) quasar absorption lines; and (3) the damping tail of the Cosmic Microwave Background temperature fluctuations. These approaches will now be discussed in turn.

Galaxies that are currently forming a new generation of stars are referred to as "star-forming galaxies", and usually have blue colors owing to the presence of massive (and therefore short-lived) stars of spectral type O and B. These hot stars produce significant quantities of photons that are capable of ionizing the atoms in the surrounding area. Electrons that have been ionized from atoms eventually recombine with another atom and produce emission lines as the electron cascades down the energy levels. An image and a spectrum of one of the most metal-poor star-forming galaxies currently known is shown in Figure 6; this galaxy is called I Zwicky 18 (where the "I" is the roman numeral for 1, and is pronounced "one Zwicky eighteen"). The ionized gas surrounding the O and B stars are generally referred to as "H II regions", because almost all of the hydrogen atoms have had their electron ionized ("H I" refers to regions that are mostly neutral, whereby most of the protons have captured an electron).

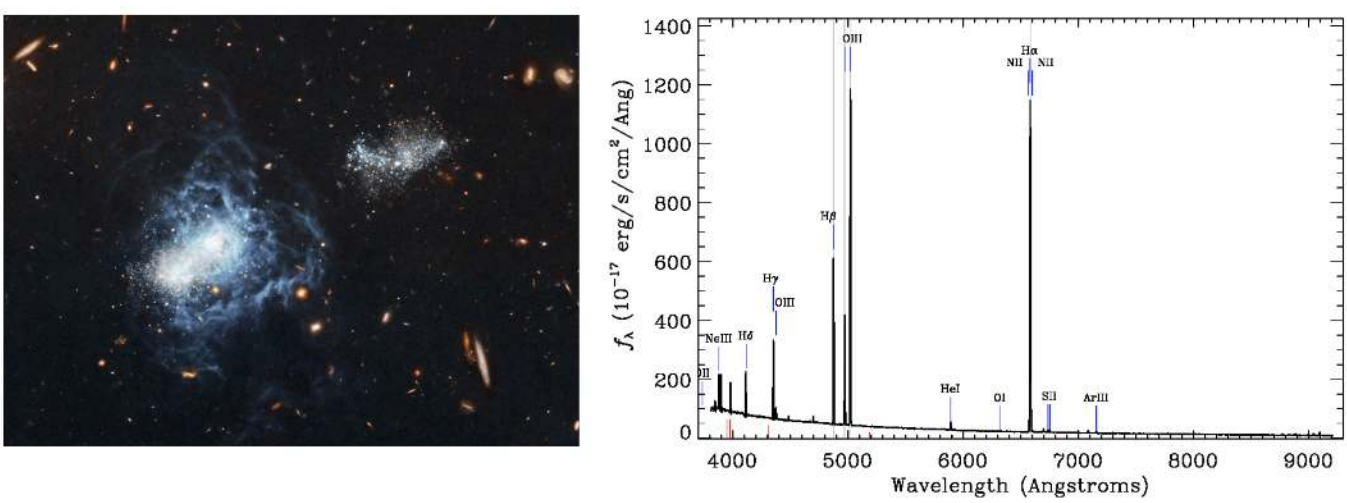


Figure 1.24.6: An image of I Zwicky 18 (left panel; credit: NASA, ESA, and A. Aloisi) and an optical spectrum (right panel; credit: SDSS). Note the blue appearance of the galaxy image and the emission lines that are detected in the spectrum, indicating that hot O

and B stars are ionizing the gas in this galaxy. The relative strengths of the emission lines can tell us about the physical and chemical properties of the gas.

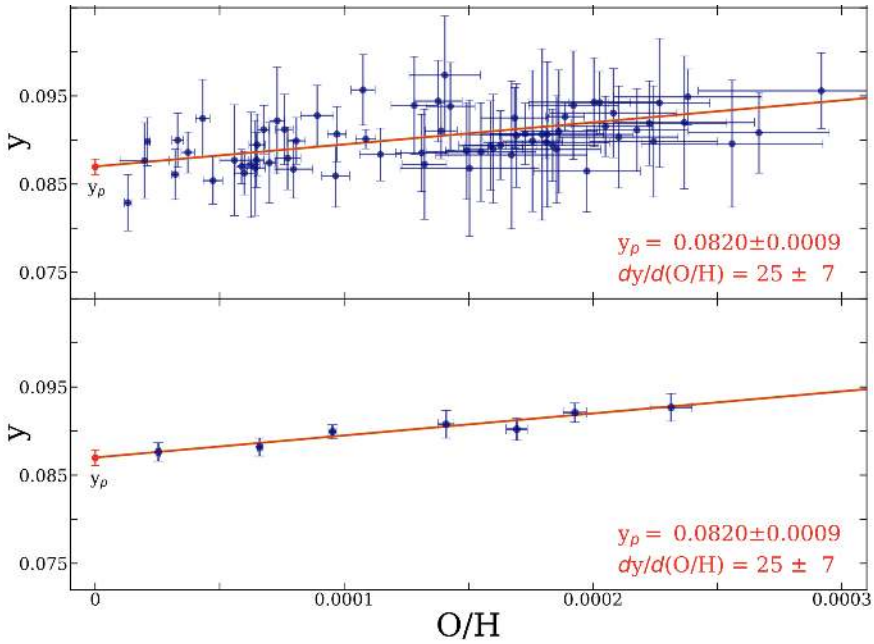


Figure $1.24.7$: The helium abundance of a sample of galaxies with different metallicities (measured as O/H) shows a gentle increase from the primordial value, y_p (credit: O. A. Kurichin).

The relative strengths of the emission lines emanating from H II regions depend on the chemistry of the gas (i.e. the relative abundance of each chemical element) and the physical conditions of the gas (e.g. the density, temperature, ionization fraction, etc.). Most of the emission lines that come from H II regions have a different dependence on density, temperature, etc. and by combining the information from many lines simultaneously, it is possible to determine both the physical and chemical properties of the gas. Measurements of the ^{4}He/H ratio in different galaxies indicate that the ^{4}He abundance gradually increases as stars produce heavy elements (see Figure 2). Therefore, the helium abundance that we measure in these star-forming galaxies does not reflect the primordial composition. Instead, we need to measure the helium abundance in many galaxies covering a range of metallicity, fit a linear relation to the helium and metal abundance, and extrapolate this linear fit to zero metallicity (see Figure 7). Current determinations of the primordial helium abundance using this approach are limited by systematic uncertainties at the ~ 1 per cent level.

The second approach that has been used to infer the primordial helium abundance uses quasar absorption line spectroscopy (see the section on Deuterium, above, for an explanation of this technique) of metal-poor gas clouds at high redshift ($z \sim 1-2$). This approach is qualitatively similar to the approach used to measure the deuterium abundance, however, there is a key difference between these approaches. For deuterium, one tries to find mostly neutral gas clouds to facilitate the detection of many deuterium Lyman series transitions. However, these mostly neutral gas clouds absorb essentially all of the background quasar photons with energies greater than the ionization potential of hydrogen (13.6 eV $\equiv 912$Å). Since all of the helium transitions occur at wavelengths < 912Å, there is insufficient quasar flux to detect helium absorption. Therefore, in order to detect both helium and hydrogen absorption lines, the gas cloud needs to be transparent to photons at wavelengths < 912Å; this occurs for gas clouds that are mostly ionized, with a column density of neutral hydrogen $N(\mathrm{HI}) < 10^{17}$ atoms cm^{-2}. The helium abundance has only been measured in one gas cloud with a metallicity of 1/30 solar, which is a similar metallicity to the most metal-poor star-forming galaxies used to measure the helium abundance. The measurement precision of this technique is currently at the ~ 10 per cent level.

The third approach that is currently employed to infer the primordial ^{4}He abundance uses the Cosmic Microwave Background temperature fluctuations (see Chapter 27: Cosmic Microwave Background Anisotropies). For further details about this technique, see the Advanced Topic at the end of this chapter.

Lithium-7 and the Cosmic Lithium Problem

The best available observational determination of the primordial ^{7}Li abundance is based on measurements of the ^{7}Li absorption from the atmosphere's of metal-poor stars in the Milky Way. This is not a straightforward measurement, because stars are extremely good at burning ^{7}Li. Convective motions (particularly in cool stars) mix the ^{7}Li near the surface of the star into the deeper layers where ^{7}Li is burned. Meanwhile ^{7}Li-deficient material is brought to the surface. To mitigate this process, observations typically focus on the hottest metal-poor stars, which have thin convective zones, and the measured ^{7}Li abundance does not correlate with temperature. This key realization came from F. Spite and M. Spite in 1982, who reported the first measurement of the ^{7}Li abundance in metal-poor stars; using only the hottest halo stars, Spite & Spite found that the ^{7}Li abundance of these stars is independent of temperature and metallicity, and has a very small scatter. This seemingly constant ^{7}Li abundance is referred to as the "Spite Plateau" (see Figure 8), and this value has remained impressively constant over the last few decades.

For many years, it was assumed that the Spite plateau was representative of the primordially produced abundance of ^{7}Li/H. However, at the turn of the millennium, the first WMAP results provided an impressively tight bound on the baryon density, and this revealed that the Spite plateau disagreed significantly with the Standard Model value, based on the CMB-derived baryon density. This problem has come to be known as the "Cosmic Lithium Problem". Many groups have re-measured the lithium abundance with increasing detailed models of stellar atmosphere's, and the result was essentially unchanged (actually this made the agreement with the Standard Model a little worse). Several nuclear physics groups searched for resonances with the reaction rates involving ^{7}Li, with the expectation that an unidentified resonance might alter the primordially predicted ^{7}Li abundance; the refined measurements of the reaction rates made the disagreement even more pronounced.

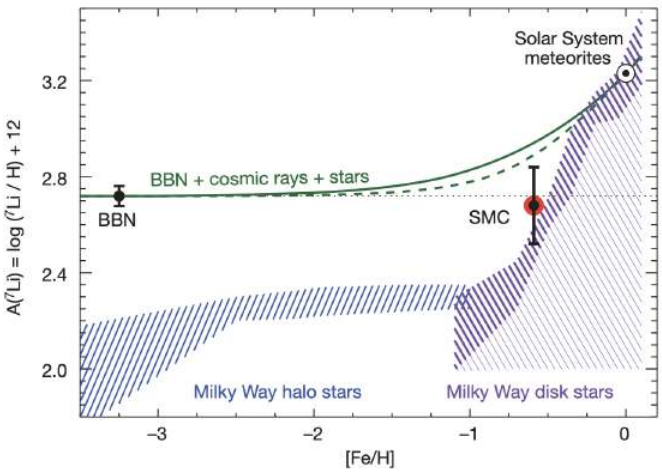


Figure 1.24.8: The chemical evolution of ^{7}Li in the Milky Way. The x-axis shows the iron abundance on a log scale relative to solar (i.e. 0 is the solar iron abundance, -1 is one-tenth solar, etc.). The primordial ^{7}Li abundance assuming the Standard Model is shown by the black symbol with error labeled "BBN". As the Universe becomes more enriched with iron, some ^{7}Li is made by stars and cosmic rays. The build-up of ^{7}Li as the iron abundance increases is shown by the green solid line, and the solar value is given by the $\odot$ symbol. The measured value in stars is shown by the purple and blue hatched region. The "Spite Plateau" refers to the roughly constant value of ^{7}Li/H between -2.5 < [Fe/H] < -1.0. At extremely low metallicities, [Fe/H] < -2.5, the Spite Plateau declines. The ^{7}Li abundance of the Small Magellanic Cloud (based on interstellar absorption line spectroscopy) is shown as the red and black symbol with an error bar (Credit: J. C. Howk).

It's fair to say that most, but far from all, of the community believe that Cosmic Lithium Problem can be solved with a better understanding of ^{7}Li burning in stars; this is currently the favored explanation for the relatively low abundance of lithium in metal-poor stars. Moreover, several groups have now measured the ^{7}Li abundance in extremely metal-poor stars, and found that the Spite Plateau appears to break down at metallicities less than 1/300th of solar (see Figure 8), giving further weight to the idea that the Cosmic Lithium Problem is a result of stellar burning of ^{7}Li. In an attempt to mitigate the problem of ^{7}Li burning in stars, a measurement of the interstellar abundance of ^{7}Li was made in the lowest metallicity environment where this measurement is currently possible - the Small Magellanic Cloud (SMC), which has an iron abundance about one-fifth of the solar metallicity. The technique used is similar to quasar absorption line spectroscopy (see description in the section on deuterium, above), but instead of a quasar, a bright O star is used as the background light source probing the interstellar medium of the galaxy. Coincidentally, this measurement is consistent with the Standard Model primordial value; however, we expect that the interstellar medium of the Small Magellanic Cloud is enriched with some ^{7}Li made by stars and cosmic rays. Subtracting the ^{7}Li produced by stars and cosmic rays in the Small Magellanic Cloud would result in a value that is consistent with the Spite Plateau, and marginally inconsistent with the Standard Model value. Thus, further work is needed to elucidate the cosmic chemical evolution of ^{7}Li, and pin down with confidence an observational determination of the primordial ^{7}Li abundance.

Helium-3

^{3}He has proven to be a very challenging primordial nuclide to measure. First, almost all ^{3}He and ^{4}He transitions are very close in wavelength (the difference is one neutron in the nucleus - see the section above on deuterium for a discussion about isotope shifts). The proximity of these transitions, and the fact that ^{4}He is 10,000 times more abundance than ^{3}He, has rendered measurements of the ^{3}He abundance to be nearly impossible. To date, ^{3}He has only been measured in the Milky Way, including a small handful of H II regions and a few environments in our Solar System (Jupiter, in particular, provides a good estimate of the pre-solar ^{3}He/^{4}He abundance). Unfortunately (for measurements of the primordial abundances), the Milky Way has experienced a significant amount of chemical enrichment, and the ^{3}He abundances measured in the Solar System and in some of the Milky Way H II regions do not reflect the primordial value. Nevertheless, despite the significant build-up of metals in the Milky Way, the latest models of Galactic chemical evolution suggest that the abundance of ^{3}He in the outskirts of galaxies like the Milky Way may be close to primordial (see the red curve in Figure 9).

The most common approach currently used to infer the primordial ^{3}He abundance utilizes the 8.7 GHz spin-flip transition of ^{3}He$^+$ (i.e. singly ionized ^{3}He). The key benefit of this approach is that ^{3}He has a non-zero nuclear spin, while the nuclear spin of ^{4}He is

zero. The ground state of ^{3}He is therefore split into two hyperfine structure levels; this splitting does not occur for ^{4}He, so the detection of ^{3}He in emission is made somewhat less difficult. Over many decades of work, there are just five reliable measurements of the ^{3}He abundance in Galactic H II regions. The most distant, and well-characterized of these H II regions is called "S209", and is 16 kpc from the Galactic centre, where models of Galactic chemical evolution predict a very modest enhancement in the amount of ^{3}He above the primordial value (see Figure 9). The value measured in this H II region agrees with the Standard Model value to within ~ 30 per cent.

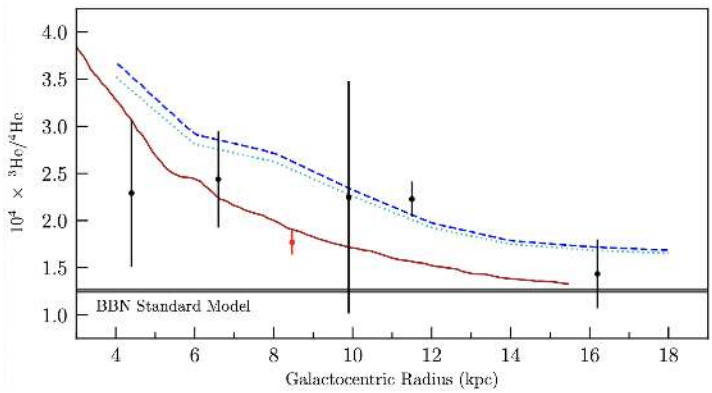


Figure 1.24.9: The present-day radial distribution of ^{3}He in the Milky Way (symbols with error bars) together with a detailed model of Galactic chemical evolution (red curve) and the primordial abundance (gray horizontal band labeled "BBN Standard Model"). The blue curves show Galactic chemical evolution models that do not include outflowing gas. The black points correspond to measurements of the 8.7 GHz fine-structure line of $^3He^+$, while the red symbol represents a measure of the ^{3}He/^{4}He isotope ratio of the Orion Nebula, using absorption line spectroscopy (Credit: R. Cooke).

An alternative approach uses the same 8.7 GHz transition to detect $^3He^+$ in absorption against the light of a radio-bright background quasar close to helium reionization at redshift $z \simeq 3.5$. While the astronomical telescope facilities do not currently exist to perform this measurement, this is an exciting goal for future facilities, because the intergalactic regions that would be probed with this technique are almost certainly of near-primordial composition.

Finally, a recent measurement of the helium isotope ratio of the Orion Nebula was reported using absorption line spectroscopy of a star in Orion to study the gas that lies at the edge of the Orion Nebula. This approach is qualitatively similar to the approach used to measure the deuterium abundance of gas clouds at high redshift (see the section on the deuterium abundance above for a discussion about this approach). Future measurements using this technique may help to precisely pin down the Galactic chemical evolution of ^{3}He, and also obtain a determination of the primordial helium isotope ratio from the outskirts of the Milky Way. It is also possible to apply this technique to gas clouds at higher redshift, where the Universe has had less time to pollute the primordial signature.

Advanced Topic - Helium-4 and the Cosmic Microwave Background

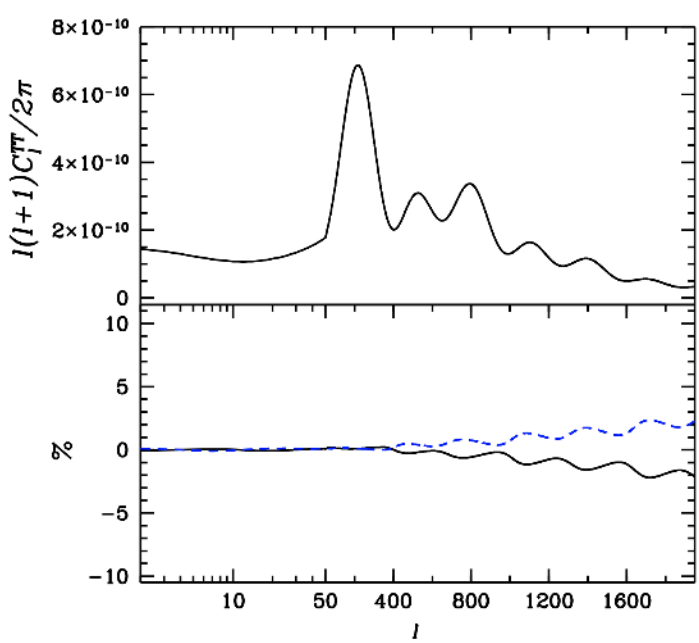


Figure 1.24.10: The power spectrum of the Cosmic Microwave Background radiation (top panel), and the deviations of the power spectrum (in per cent) due to changes of the helium abundance. The blue dashed and solid black lines represent a decrease and increase of the primordial helium abundance by 10 per cent. (credit: Trotta & Hansen 2003).

An alternative and very promising approach to measure the helium abundance utilizes the power spectrum of temperature fluctuations imprinted on the Cosmic Microwave Background radiation. This approach has the advantage that helium is purely primordial during recombination, because there are no channels (i.e. stars) to produce/destroy ^{4}He during the first 400,000 years after Big Bang nucleosynthesis.

The key physical process that is sensitive to the helium abundance is diffusion damping. Note also that if more ^{4}He is produced during Big Bang nucleosynthesis, that means less hydrogen is made, and vice-versa. Diffusion damping (sometimes referred to as Silk damping) is a process whereby the temperature fluctuations are smoothed out on small scales (high

multipoles). The damping is a result of photons being constantly scattered by charged particles (mostly electrons) prior to recombination. Photon scattering stops when the temperature of the Universe drops to the point that electrons can recombine with nuclei, at which point photons can free stream. Electrons recombine with helium first, followed later by hydrogen recombination; this is because helium has a higher ionization potential than hydrogen (i.e. it is more difficult to ionize an electron from helium once it has recombined). If more helium is made during Big Bang nucleosynthesis, there are fewer electrons between helium and hydrogen recombination, leading to less scattering. This allows photons to free stream further, damping the perturbations and reducing the CMB power at small scales. The effect is relatively small (see Figure 10), and the current inference on the primordial helium abundance using this approach is ~ 10 per cent. Future CMB experiments that target small angular scales aim to measure this quantity to much higher precision.

This page titled 1.24: Big Bang Nucleosynthesis - Observations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Ryan Cooke.

1.25: Introduction to the Cosmic Microwave Background

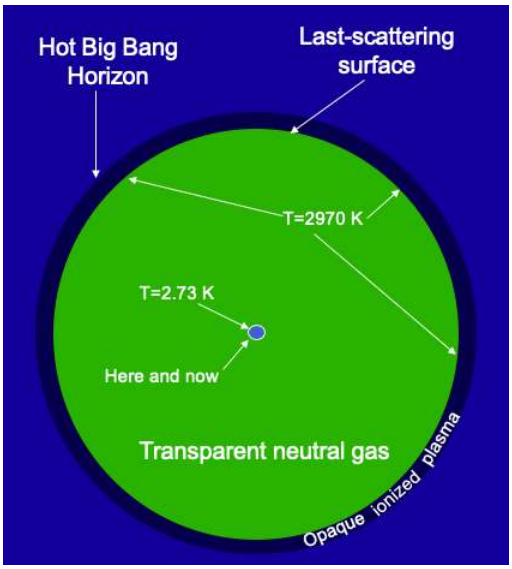
Discovery

As we have seen in our study of big bang nucleosynthesis, Alpher and Herman predicted the cosmic microwave background (CMB) in the 1940s. But it was decades until anyone mounted a serious campaign to look for this relic of the big bang. The delay seems astounding from our perspective, and yet nonetheless this is what happened. The prediction, of a relic of the big bang, that would not have been too difficult to detect, somehow did not garner much attention. The first effort to make measurements to detect the CMB was launched by a group at Princeton University in the early sixties. They had an extremely well-motivated set of observations in mind. They had the technical ability to build the instruments and make the measurements. They had the funding they needed. They just got started a little too late: in 1964 they were scooped by a pair of radio astronomers, a mere 30 miles to their east at Bell Labs, who had stumbled upon the cosmic relic accidentally. Life is not always fair.

Arno Penzias and Robert Wilson were the Bell Labs-employed radio astronomers, who, though they did not know it, were performing measurements of great historic importance. They had re-purposed a radio telescope, built for early experiments with satellite communication, to do some astronomy, such as measuring radio emission from a supernova remnant. As part of a process of characterizing their instrument and its noise properties, they made measurements away from any such sources. They were unable to reconcile the output from their radiometer, when the telescope was aimed at blank sky, with their estimates of the amount of noise there should be in their measurements. No matter where they pointed their telescope, the radiometer output exceeded what they expected from noise, consistent with a uniform background of microwaves.

The two groups became aware of each other's efforts (via some intermediaries), which led to them all realizing that the CMB had been detected. They agreed to write a pair of papers. The first was authored by Penzias and Wilson and titled "**A Measurement of Excess Antenna Temperature at 4080 Mc/s.**" The second was written by the Princeton group and provided the cosmological interpretation of their measurement, simply titled "Cosmic Black-body Radiation." Penzias and Wilson received the 1978 Nobel Prize in Physics for their discovery.

The Last-Scattering Surface



Before going much further it's worth having a picture in your mind of what we're looking at when we're looking at the CMB. Toward that end, consider the figure on the right, a cartoon of our past light cone. To see it as a cone, you'll have to use your imagination to lift the "Here and now" part up off the screen to be the tip of the cone, and then have the circle labeled as "Hot Big Bang Horizon" remain back down on the surface of the screen forming the base of the cone. The axis coming out of the screen is the temporal dimension. The center is our location in spacetime (where we are in space and when we are). Moving radially away from the center, one is moving away from us in comoving distance and, because we are staying on the cone, also moving back in time.

As we have already seen, if we assume a radiation-dominated universe we find that it reaches infinite density a finite time in the past. Further, even allowing for the transition from radiation domination to matter domination, and then to cosmological constant domination, a signal traveling at the speed of light from that time of infinite density until today will only travel a finite distance. We call that distance the past horizon. That distance is the radius of the circle labeled "Hot Big Bang Horizon."

But light can't travel stright to us from some time way back in the radiation-dominated era because the universe is filled with an opaque plasma. A plasma is the state of matter when it is sufficiently hot that the electrons are stripped away from the nuclei. Since light readily scatters off of free electrons, plasmas at sufficient density and sufficiently thick are opaque; light can't travel freely through them. However, as the universe expands, it cools, and the universe undergoes a transition from an opaque ionized plasma to a transparent neutral gas. We've labeled this transition on our diagram as the "Last-scattering surface" because this is where light that arrives here and now last scatters off of a free electron. We also refer to this epoch as "decoupling" because it is when light decouples from matter.

So, when we are detecting CMB photons, we are detecting light that, for the most part, last interacted with matter at the time of this transition. If our eyes were sensitive to light at microwave frequencies, when we looked up at the sky we would be seeing an image of this last-scattering surface. Now in the diagram the last-scattering "surface" is just a circle. But that's because, in order to show this diagram, we have suppressed one spatial dimension. Instead of a circle, the last-scattering surface is actually a sphere. It's the surface around us, just the right distance away that photons scattering off of electrons for the last time, headed our way, are just getting here now.

A (Somewhat Boring) Map of the CMB

Penzias and Wilson had unwittingly stumbled upon the CMB. A key property about it they noticed, before they even knew it was the CMB, was its isotropy; no matter what seemingly empty piece of sky they pointed their telescope towards they got about the same amount of signal (or 'excess noise' as it appeared to them). We now know that it is isotropic to about one part in 1,000. This is such a high-degree of uniformity that a full-sky map of the CMB intensity looks completely uniform. The full-sky map below was made with data from the Differential Microwave Radiometer (DMR), an instrument launched on the COsmic Background Explorer (COBE) satellite in 1989. The high degree of uniformity of this map reflects the high degree of uniformity of the universe at the time of decoupling.

We refer to a quantity averaged over the whole sphere as a 'monopole.' Things get a tiny bit more interesting when we subtract this average intensity and look at what remains.

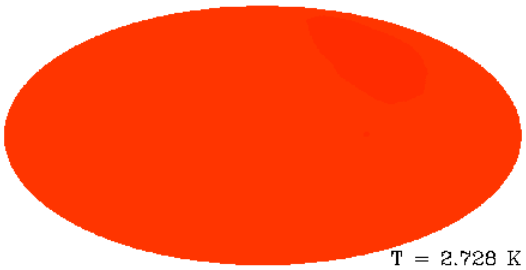


Figure 1.25.2: A full-sky map of the CMB inferred from data from the COBE satellite. The lack of variation in color indicates the high degree of isotropy of the CMB; it is nearly the same brightness (and temperature) in all directions.

The CMB Dipole

The Earth moves around the Sun, which orbits the center of the Milky Way. The Milky Way galaxy itself is falling towards Andromeda and the Local Group of galaxies is falling toward the Virgo Cluster. It is no surprise then that we are moving with respect to the rest frame of the cosmic microwave background. Our relative motion induces a dipole pattern on the sky, as can be seen in the map below which has had the monopole subtracted from it. We see a dipole pattern, as well as contamination by some emission from our own galaxy clustered near a horizontal line that goes through the center of the map. Radiation in the Earth's direction of motion appears blueshifted and hence hotter, while radiation on the opposite side of the sky is redshifted and colder.

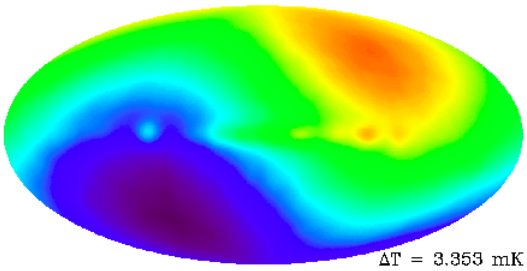


Figure 1.25.3: The CMB as mapped by the COBE satellite, after monopole subtraction. The dipole component is clearly visible. One can also see some emission from the plane of our own galaxy.

Beyond the Monopole and Dipole

With the monopole and dipole subtracted, the remaining variations in the temperature of the CMB are at the level of tens of microKelvin, reflecting small variations in conditions on the last-scattering surface. We will see such maps of the CMB in subsequent chapters and learn how we have used them to probe the dynamics of the primordial plasma. Before that though we will study the spectrum of the CMB, nature's best approximation to black body radiation.

Box 1.25.2

Exercise 25.1.1: Provide a 6 to 10-sentence summary of this chapter.

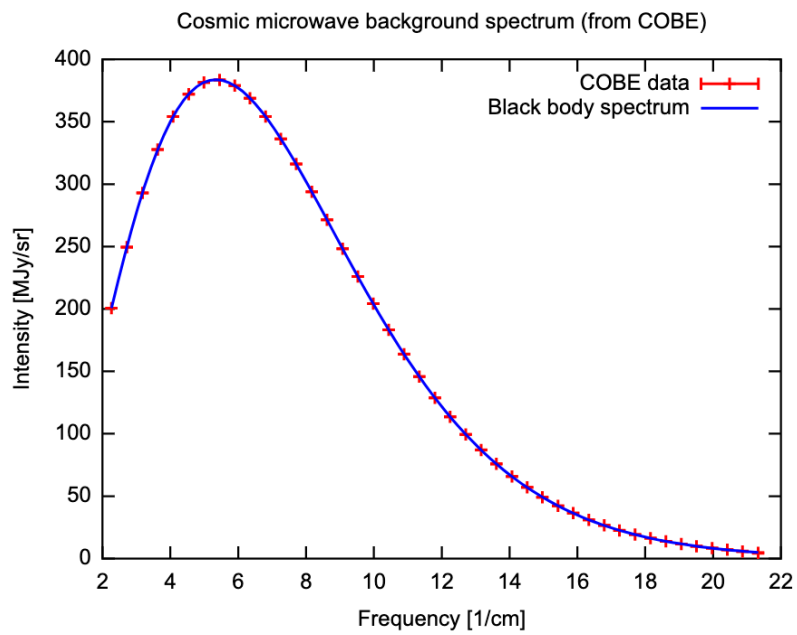
This page titled 1.25: Introduction to the Cosmic Microwave Background is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.26: The Spectrum of the CMB

No other natural source of radiation has ever been measured to be as consistent with black-body radiation as is the case with the CMB. Here we look at the measurements of the spectrum from a Nobel-prize winning instrument on the COBE satellite, before turning to the question of *why* the CMB is so near to being a black body and what we can learn from that. We will see that this is expected based on a very early epoch in which a thermal distribution is established by kinetic-energy-changing, and photon-number-changing interactions, and subsequent epochs in which passive evolution under the expansion of space preserves the thermal distribution achieved earlier, albeit with a temperature that decreases as the universe expands. The very high number of photons relative to matter also plays a role in making the CMB a better approximation of a black body than is the case for the Sun, for example.

The Spectrum Observed

NASA's COsmic Background Explorer (COBE) satellite was launched in 1989 with three scientific instruments on board, two of which would lead to Nobel prize-worthy discoveries. One of these two was the Far Infrared Absolute Spectrometer, or FIRAS. The precision and accuracy of the FIRAS measurements were extraordinary. The figure below shows the spectrum of the CMB as determined from FIRAS data. The measurements are so precise that the width of the error bars is smaller than the thickness of the theoretical black body spectrum curve. You can read the original paper describing the FIRAS measurements and determination of this spectrum here: https://iopscience.iop.org/article/10.1086/178173/pdf.



The "Intensity" plotted on the y axis is telling us the energy per unit amount of sky (solid angle) per unit interval of frequency flowing through a unit area per unit time. The units are MJy/sr where "Jy" stands for Jansky with 1 Jansky = 10^{-23} erg/s/cm^2/Hz and "sr" is short for steradian, a unit of solid angle. The x axis is the frequency in units of 1/cm -- which is an unusual way of expressing frequency. To convert to more familiar units, one multiplies by the speed of light, which is about 3×10^{10} cm/sec so, for example, 10 cm^{-1} is about 300 GHz.

Theoretical Prediction

According to our standard thermal history, reactions that create and destroy photons were fast enough at $z \geq 10^7$ to drive the CMB photons into a black body distribution; i.e., a thermal distribution with zero chemical potential (for massless bosons). In terms of the phase space distribution function, f we have

$$f = \frac{1}{\exp(pc/k_B T) - 1}. \tag{1.26.1}$$

So that's true for $z \geq 10^7$. What about today? We need to work that out in order to compare with observations we can do today. At $z \leq 10^7$ the number changing reactions are slow so we are not guaranteed that the photon chemical potential remains zero. Let's work out what happens under the assumption that no new energy is injected into the photon distribution (by an unstable particle that has decay byproducts that include photons, for example). Let's do so by tracking what happens to photons in a small region of momentum space initially centered on momentum $\vec{p}_i$ with volume $V_{p,i}$, and in some comoving volume, with initial physical volume $V_{x,i}$. We've taken the phase space volume to be small enough that f does not vary by much across it. So we don't need to integrate to find the total number of particles, we can just multiply:

$$N_i = f(p_i; T_i) V_{x,i} V_{p,i}. \tag{1.26.2}$$

Let's, for now, ignore scattering interactions that the photons have and treat them as if they are not receiving any kinetic energy from scattering. In this case their momenta are just redshifting with the expansion so that at some later point in evolution, when the scale factor has gone from a_i to a_f, we have $p_f = p_i a_i / a_f$. Also due to the redshifting of momentum, the momentum-space volume that contains these photons shrinks by a factor of $(a_i/a_f)^3$, and the physical volume of the comoving volume we are tracking grows by $(a_f/a_i)^3$. So we have

$$N_f = f(p_f; T_i) V_{x,f} V_{p,f} \tag{1.26.3}$$

$$= f(p_f; T_i) \left(\frac{a_f}{a_i}\right)^3 V_{x,i} \left(\frac{a_i}{a_f}\right)^3 V_{p,i} \tag{1.26.4}$$

$$= f(p_f; T_i) V_{x,i} V_{p,i}. \tag{1.26.5}$$

Box 1.26.1

Exercise 28.1.1: Draw the Cartesian axes of a 2-dimensional momentum space (p_x, p_y) and also of a 2-dimensional configuration space (x, y). Draw a box centered on the origin of the configuration space and one around some region of the momentum space, representing boxes bounding the particles we are going to track from one time to another. Take this first drawing as for the initial time, and make a subsequent pair of graphs for the later time, with the sizes and locations of these boxes evolving appropriately, following the above discussion. Take the configuration space coordinates to represent physical distances of particles from the origin; i.e., these are not comoving coordinates.

How is N_f related to N_i? By design we've tried to keep these two equal. We've shifted the region of momentum space to track the photons as they redshift. So none of them have moved out of our momentum-space region due to how the momentum has changed. They are moving through space in different directions, so some will leave our comoving spatial volume. But, due to homogeneity and isotropy, the same number will flow in as flow out. The net result is that $N_f = N_i$. From this we can conclude that

$$f(p_f; T_i, t_f) = f(p_i; T_i, t_i) = \frac{1}{\exp(p_i c / k_B T_i) - 1} \tag{1.26.6}$$

$$= \frac{1}{\exp(p_f (a_f/a_i) c / k_B T_i) - 1} \tag{1.26.7}$$

$$= \frac{1}{\exp(p_f c / (k_B T_i a_i / a_f)) - 1}$$

$$= \frac{1}{\exp(p_f c / k_B T_f) - 1} \tag{1.26.8}$$

with $T_f = T_i \frac{a_i}{a_f}$.

This is a remarkable result. Even though we ignored all interactions, the things that maintain equilibrium, we find out that such evolution results in a phase space distribution function that is the same as one would get for thermal equilibrium with zero chemical potential, although at a reduced temperature. The expansion of the universe cools the gas of photons while maintaining a chemical-potential-free thermal distribution.

We have ignored any energy that might be imparted to the photons by the electrons. But any such energy is very small since, in a given region of space, the total kinetic energy in the electrons is much less than the total kinetic energy in the photons, mainly because of how the photons greatly outnumber the electrons. The bottom line is that we expect, to a good approximation, the photons today to have a phase space distribution function for massless bosons in thermal equilibrium with zero chemical potential. Given that they have two degrees of freedom, this is a Planck distribution.

This page titled 1.26: The Spectrum of the CMB is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.27: Cosmic Microwave Background Anisotropies

Introduction

The plasma that existed from the first fractions of a second of the Big Bang until it transitioned to a neutral gas 380,000 years later is a beautiful gift of nature. Gently disturbed away from equilibrium by mysterious, very early-universe processes, the plasma is an unusually simple, natural dynamical system. Due to its simplicity, we can calculate the evolution of these disturbances with high accuracy and use these calculations to predict observable consequences. The agreement between our predictions and precision observations is extraordinary, allowing us a high degree of confidence that we understand in detail events that were transpiring 14 billion years ago, throughout the cosmos, as far away as the edge of the observable universe.

The simplicity of this system is perhaps best understood in contrast to other natural systems. Naturally occurring phenomena are usually too complicated to understand fully from first principles. We know the underlying physics, but the equations are really complicated to solve, and we'd need an absurd amount of data to narrow down from the set of all possible solutions of the equations, to the one that is actually being realized with the phenomenon in question. Examples of complicated phenomena are supernova explosions, earthquakes, and human behavior. We think we know the underlying physics at play, but getting from there to predictions for observations is extremely difficult, if not impossible.

Usually, to observe a system for which we can make predictions, we need to carefully construct the system ourselves. We call such carefully constructed scenarios, 'experiments.' Although the exception rather than the rule, there *are* examples of naturally occurring systems that are amenable to our understanding. The solar system is such a naturally occurring system, with its simplicity arising from the fact that gravity is the dominantly important force, and the high accuracy of the simplifying approximation of the planets and stars as point masses. The simplicity of this system, and the regularity of the passage of the planets and Sun across the sky may have, in fact, played an important role in the discovery of science itself, as argued by Steven Weinberg in *To Explain the World: The Discovery of Modern Scienc*e.

Another such simple system is this early plasma, often called the primordial plasma. We are motivated to share with you our observations of the sky at millimeter to sub-millimeter wavelengths (observations that reveal patterns in this plasma just as it disappeared about 14 billion years ago) and the remarkable agreement of the statistical properties of these observations with our calculations. To fully appreciate this agreement, we will introduce you to some fundamental aspects of how those calculations are done.

In this chapter, we will introduce the primordial plasma and how observing the sky in millimeter to sub-millimeter wavelengths allows us to see the plasma just as it disappears by transitioning to a neutral gas. In this section we will also see the extraordinary agreement between the statistical properties of maps of the CMB and our predictions for these statistical properties, predictions that follow from our understanding of the dynamics of the plasma. The main focus of this chapter is to gain a qualitative understanding of a statistical property of CMB maps called a 'power spectrum.' In the next chapter, we will model the plasma as a fluid and introduce a simple version of the dynamical equation that governs the evolution of its density, a wave equation. We will find solutions to the wave equation, and demonstrate the benefits of Fourier decomposition for evolving a system governed by such an equation. Then we will work out the prediction of the main qualitative feature of the spectrum -- the existence of a series of peaks.

Note

Before we fully begin, let me digress with one bit of history here related to the prediction of this series of peaks and their detection. On the left is a photo I took at the June 30, 2001 launch of the Wilkinson Microwave Anisotropy Probe (WMAP). From left to right are four fellow theorists: Ned Wright, Neil Cornish, Dick Bond, and Rashid Sunyaev. Despite Rashid's significant contributions to the field many of us, including me, had not met him before. Dick Bond was ever-present at Rashid's side, introducing him to people saying, "This is Rashid Sunyaev, as in Sunyaev-Zel'dovich" like he was bringing you in on some funny joke, as everyone in my field knows of something we call the Sunyaev-Zel'dovich effect. One couldn't help but smile in response. Rashid, meanwhile, was quite emotional, telling me that his advisor (Zel'dovich) had told him, "This is *beautiful* physics, but it will never be observed. Work on something else now." I was so ignorant of the history, I only realized later he was talking about

their 1967 work that was the first to point out that there would be acoustic peaks in the power spectrum of the CMB. We've now measured those peaks with very high precision. The WMAP satellite launched that day worked spectacularly well and was a big step forward in our measurements of the CMB. The launch was a joyful occasion.

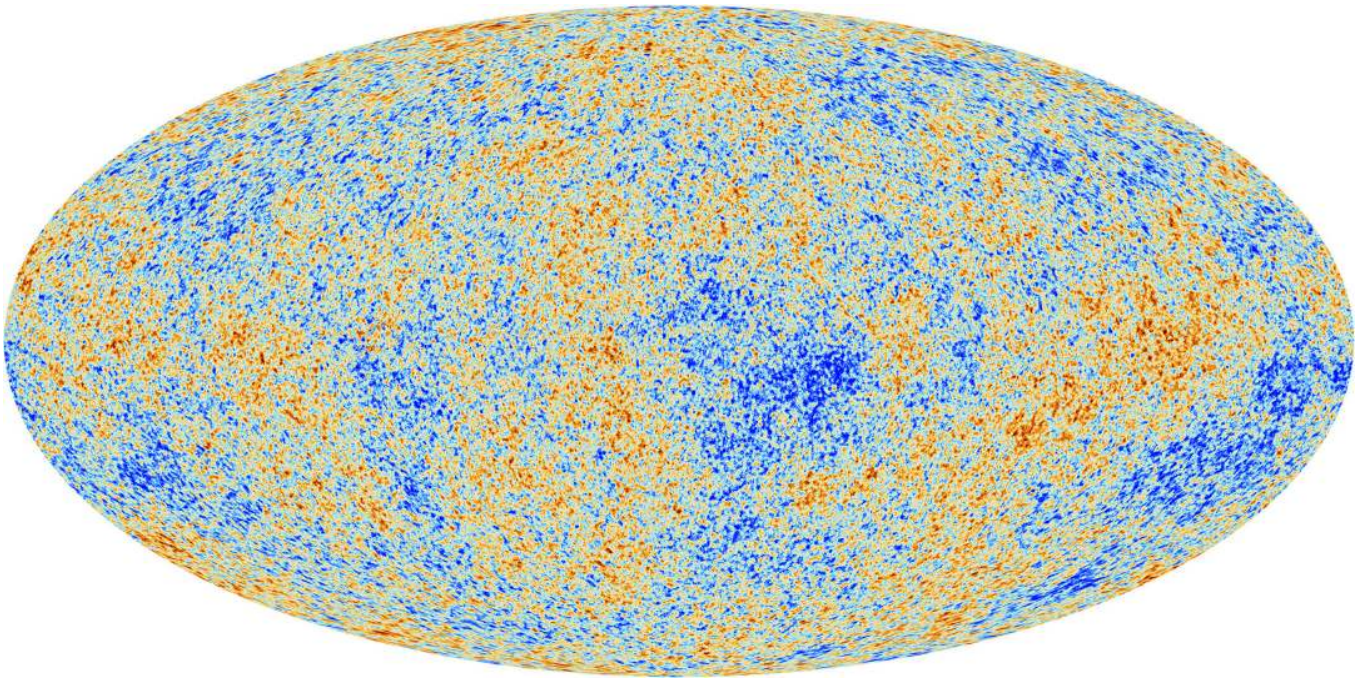
The Primordial Plasma

A plasma is a state of matter in which electrons are dissociated from nuclei. For example, the Sun is a great ball of gravitationally-bound plasma. Plasmas are hot, as it is the thermal energy that keeps the negatively charged electrons from binding with the positively charged nuclei. Plasmas inevitably include photons, since the charged particles (protons and electrons) accelerate in the electric fields of other charged particles and radiate photons as a result. The primordial plasma was a collection of electrons, hydrogen and helium nuclei, trace amounts of other light nuclei, and photons. It existed before there were any stars, and so before there were any heavier elements that eventually were formed in stars. The formation of helium and other light elements is the subject of the classic popular book by Steven Weinberg, *The First Few Minutes*, to which we allude with the title of this chapter.

As we have discussed before, the plasma eventually cools sufficiently that the electrons bind with the protons and helium nuclei, so that the plasma transforms into a neutral gas. The universe becomes transparent and the thermal photons start freely streaming across the universe. Those photons that originated at just the right distance from us, and were headed our way, are arriving now. They give us an image of what the universe was like at the time of this transition in a thin spherical shell around us at a distance of about 46 billion light years that we call the last-scattering surface. Thus we have a means of studying the primordial plasma observationally. We saw maps of the CMB radiation in the previous chapter that revealed the monopole and the dipole. If we remove the monopole and the dipole then the remaining map is dominated by features that correspond to inhomogeneities on this last-scattering surface. A full-sky map of the CMB, as determined by data from the Planck satellite launched in 2009, is shown below. Planck flew later than WMAP but had higher angular resolution, increased sensitivity, and mapped the sky over a broader range of frequencies.

The following image is a projection of the spherical sky onto a 2D map. To better understand how this relates to what we observe from Earth, explore the virtual CMB planetarium below it by clicking and dragging. This shows how the CMB would look over a tree-lined horizon, if human eyes were extremely sensitive to light at millimeter wavelengths.

'



The CMB Power Spectrum

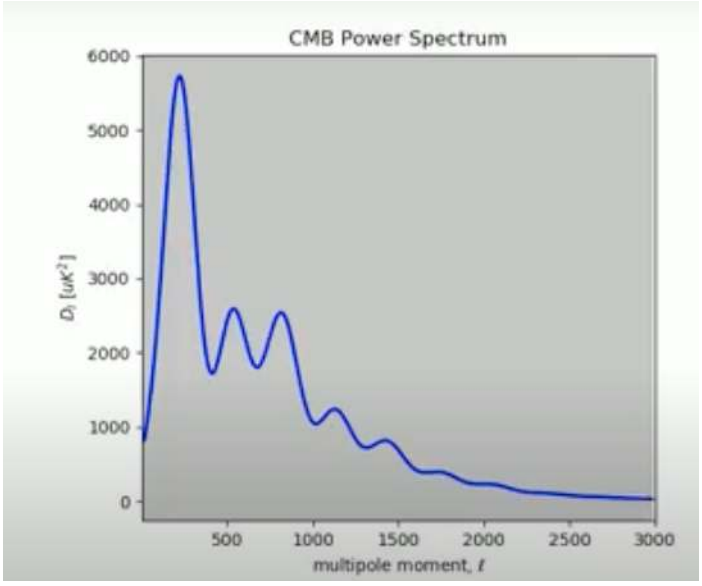
So how do we learn things from a map of the CMB? Our theories do not actually predict the map. They predict *statistical properties* of the map. The most important statistical property is called the "power spectrum." The power spectrum tells us about the smoothness/roughness of the map, as a function of angular scale.

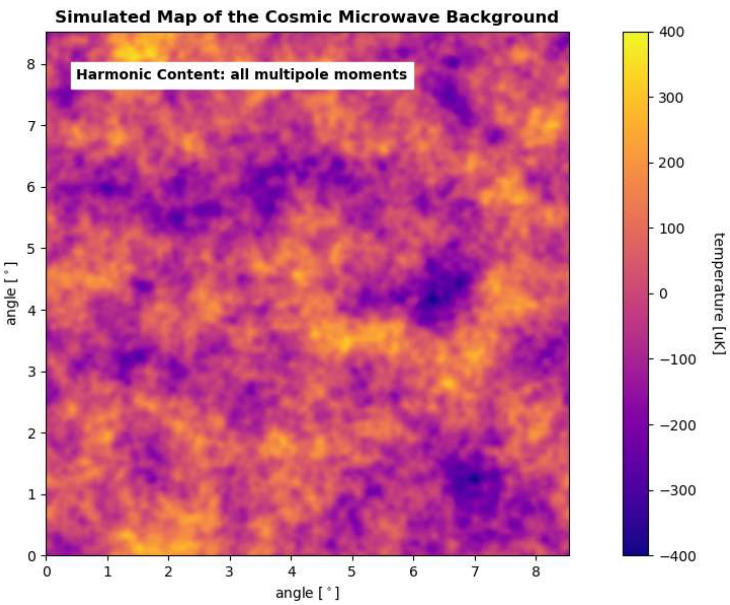
Qualitative Description

Now what does that mean, "as a function of angular scale?" Well, how would you describe the surface of the Pacific Ocean? Is it smooth or rough? The answer depends on length scale. On very large scales, scales much larger than the typical wavelength of a wave, it is quite smooth. But then on scales the length of a typical wave it is rougher. Zooming in further, to length scales smaller than a typical wave wavelength, it might appear smooth again.

Here, we'd also like to make very clear that we are *not* talking about the wavelengths of radiation emitted from the CMB. *The power spectrum of the CMB refers to spatial wavelength of fluctuations in the temperature of the CMB across the sky.*

We can do the same type of analysis with a map of the cosmic microwave background that we did with the Pacific Ocean. The scale-dependent measure of "roughness" we call "power." The power spectrum of the CMB is shown in the left panel below. The power is on the y-axis and the angular scale is shown on the x-axis. A higher multipole moment ℓ corresponds to a smaller scale. On the right panel is a simulated map that is consistent with this power spectrum. (Note that in all the following animations and pictures, each square shows an 8.5 x 8.5 degree patch of the CMB).





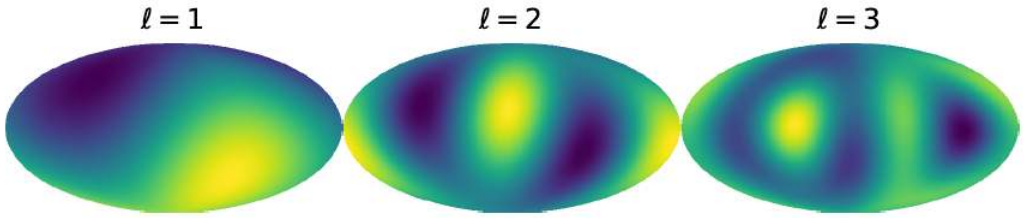
The y axis is power and the x axis is a measure of angular scale with larger scales on the left and smaller scales on the right.

Quantitative Description

Now let's give a more mathematical description of the power spectrum. Any function on the surface of a sphere, $T(\theta, \phi)$ can be written as a sum over complex functions on the sphere with well-defined wavelengths, the spherical harmonics $Y_{lm}(\theta, \phi)$:

$$T(\theta, \phi) = \sum_{lm} a_{lm} Y_{lm}(\theta, \phi). \tag{1.27.1}$$

The sum over l ranges from 0 to infinity and the sum over m ranges from $-l$ to $+l$. The l value tells you the wavelength of the oscillations: $Y_{lm}(\theta, \phi)$ undergoes l oscillations as one runs through all 360 degrees of a great circle. The spherical harmonics for the lowest values of l have names: monopole, dipole, quadrupole and octupole are for $l = 0, 1, 2, 3$ respectively. The figure below shows examples of dipole, quadrupole and octupole patterns over the whole sky (using the Mollweide projection for mapping the sphere on to a plane). Notice how larger multipoles like $\ell = 3$ correspond to smaller features on the CMB map, whereas lower multipoles like $\ell = 1$ correspond to very large scale features.



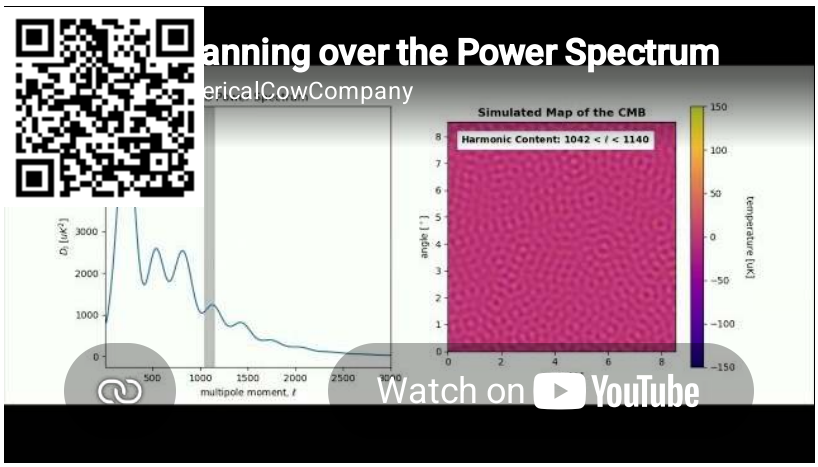
Our cosmological models do not predict the temperature in a particular direction, or the value of a particular a_{lm}; they predict statistical properties such as the mean and variance of a_{lm}. For isotropic cosmologies, the mean of all except for the monopole is zero. The variance in isotropic theories is independent of m and given by

$$C_l \equiv \langle a_{lm} a_{lm}^* \rangle. \tag{1.27.2}$$

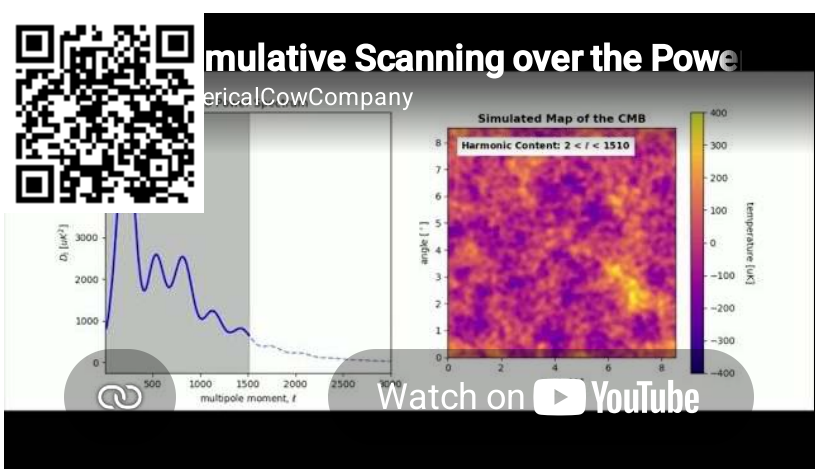
This variance, as a function of l, is called the power spectrum. Rather than C_l, we usually plot $D_l \equiv l(l+1)C_l/(2\pi)$ instead.

More Intuition about the Power Spectrum

Below, we show the same power spectrum and CMB map from above, but only features of the map whose scale falls within the grey box are included. At first, we see large scale features that increase in amplitude as the grey box passes over the first peak in the power spectrum at ℓ of about 200. We then see progressively smaller features of lower amplitude as the box scans through the higher-ℓ region. The pitch of the audio corresponds to the size of the features and the volume to the amplitude of the power spectrum in that range. We can hear the pitch increasing as we move from left to right. The volume rises as we pass over the first peak, then tapers off.

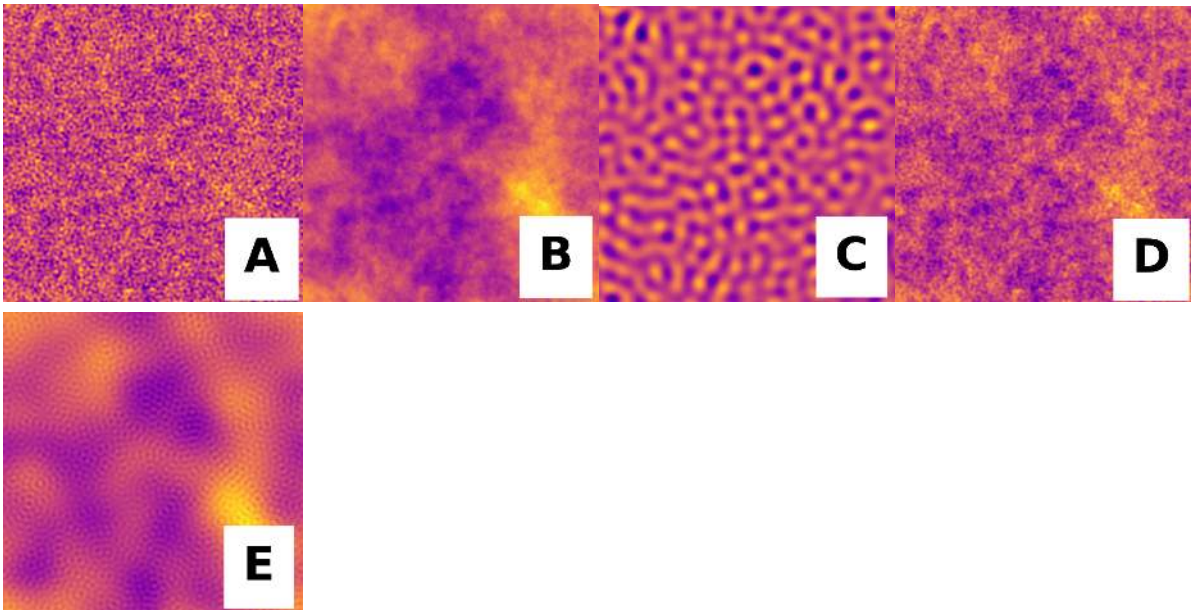


In the video below, the grey box steadily includes more of the power spectrum over time. As the box expands, we can hear more frequencies and see more features on the map.



Play "Match the Power Spectrum to the map"

First, observe each simulated CMB map. Next, click and drag to match the letter of the CMB map to the correct power spectrum on the left. Remember, large features correspond to power on the left of the power spectrum and small features correspond to power on the right side of the power spectrum. Feel free to refer back to the above videos. Warning: this is challenging. You will have to put some thought into it, and even then are likely to get it wrong the first time. If you do, think again and try again. I incorrectly assigned two of the power spectra the first time I tried it!

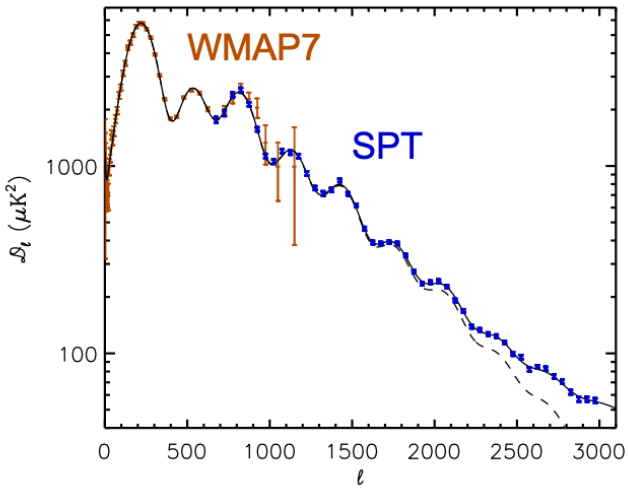


Comparison with Observations

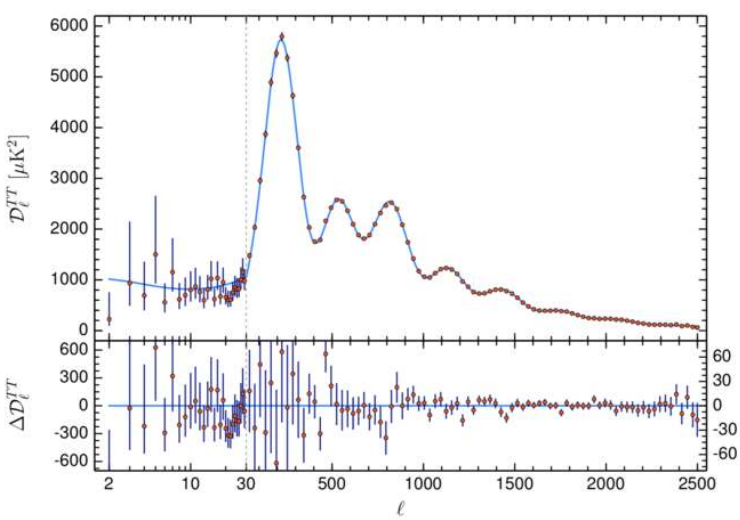
Observations of the microwave background intensity as a function of direction on the sky have been carried out since the late sixties from telescopes on the ground, on high-altitude balloons, and from three different spacecraft (COBE, WMAP, and Planck). It is useful to get above the atmosphere as the atmosphere itself emits and absorbs at millimeter to submillimeter wavelengths, so balloons and spacecraft offer advantages. Ground-based instruments have been competitive though, in particular at higher angular resolution (higher ℓ), which requires larger telescopes that are harder to lift above the atmosphere. One of the best ground-based sites for observing at millimeter to submillimeter wavelengths is at the South Pole, where the 10-meter South Pole Telescope (SPT) has operated since 2007.

The WMAP satellite, like Planck, observed the full sky, whereas the SPT survey covered 2,500 square degrees, about 6% of the sky, but with lower noise and higher angular resolution. The figure below is from Story et al. (2013), showing the determination of the CMB temperature power spectrum from 7 years of WMAP observations and from the complete SPT 2500 sq. degree survey. The solid line is a fit to these spectra that includes the contribution from the CMB itself assuming the best-fit standard

cosmological model (dashed line), and the contribution from other extragalactic sources of radiation.



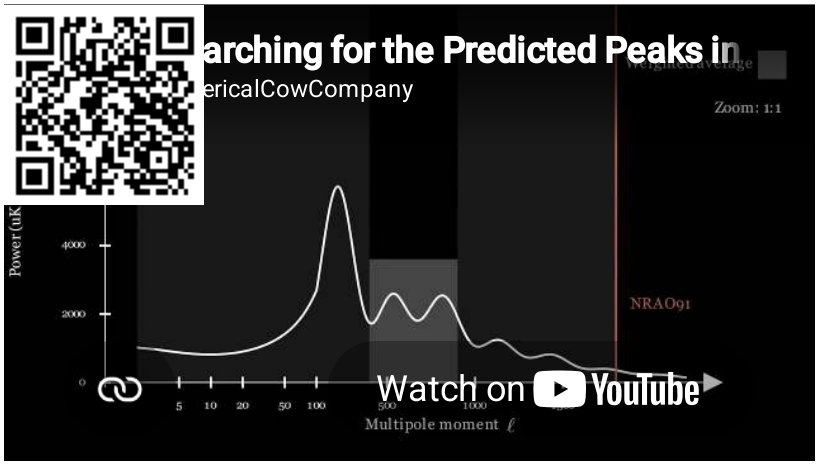
The figure below is from the Planck Collaboration showing their determination of the CMB power spectrum and the predicted power spectrum of the standard cosmological model with its parameters adjusted to best agree with the data. The lower panel shows the 'residuals': the data points after the best-fit model has been subtracted from it. It allows one to better inspect the quality of the agreement between theory and data. The small amplitude (relative to the error bars) of the departures from zero in the lower panel indicate agreement between theory and observation. Note that the y axis for the residuals is different for $\ell < 30$ than it is for $\ell > 30$.



The agreement between observation and theory here is quite extraordinary. It gives us a very high level of confidence that we have a good understanding of the physical conditions of the universe when the scale factor was over 1,000 times smaller than it is today, around 14 billion years ago.

The Progression of CMB Power Spectrum Experiments

The following video tracks the progress of CMB measurements from 1967 to 2020. Gray boxes show the weighted average of prior measurements. The error bars indicate the type and quality of a measurement. The horizontal bar indicates the range of multipoles ℓ over which the power measurement was taken. The vertical bar indicates the 1σ error. (That means there is a 68% probability the true power lies within the vertical bar). If the error bar has a downward pointing arrow, it indicates that the experiment only determined an *upper bound* on the CMB power, not a detection.



This page titled 1.27: Cosmic Microwave Background Anisotropies is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.28: Solving the Wave Equation with Fourier Transforms

[** This chapter is under construction **]

In the next chapter we will introduce the wave equation due to its importance in understanding the dynamics of the primordial plasma. In one dimension the wave equation can be written as

$$\frac{\partial^2 \Psi(x,t)}{\partial t^2} = v^2 \frac{\partial^2 \Psi(x,t)}{\partial x^2}. \tag{1.28.1}$$

We will leave a discussion of the physics of this equation and the primordial plasma to the next chapter. Here, we will focus on the use of Fourier methods to solve for the evolution of $\Psi(x,t)$ assuming it obeys the above equation and that we are given the value of Ψ and its time derivative at some initial time for all values of x. Fourier methods have a broad range of applications in physics. They have utility well beyond the dynamics of the wave equation in both experimental and theoretical physics. For the student of physics, time spent developing facility with Fourier transforms is time well spent.

Let's see what happens with an ansatz of the form

$$\Psi(x,t) = A(t)\cos(kx); \tag{1.28.2}$$

i.e., let's assume the wave has a fixed spatial pattern of a cosine of wavelength $\lambda/(2\pi)$, with an amplitude that varies with time.

Plugging this ansatz in to Eq. 1.28.1 we find that it is a solution of Eq. 1.28.1 as long as

$$\ddot{A}(t) = -v^2 k^2 A(t); \tag{1.28.3}$$

i.e., as long as $A(t)$ obeys a harmonic oscillator equation.

Box: do the above plugging in to arrive at Eq. 1.28.3.

? Exercise 1.28.1

Do the above "plugging in" to arrive at Eq. 1.28.3

Answer

TBD

The general solution to Eq. 1.28.3 is

$$A(t) = \alpha\cos(kvt) + \beta\sin(kvt) \tag{1.28.4}$$

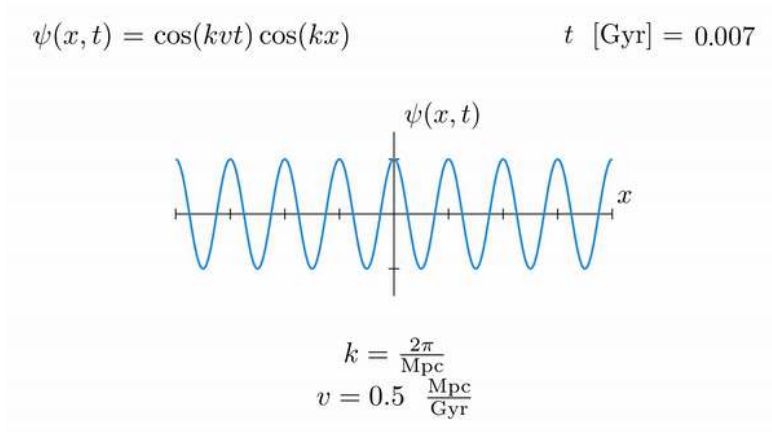
.

The constants α and β can be determined from initial conditions $A(0)$ and $\dot{A}(0)$.

Because it will be helpful to see a specific solution, let's assume the ansatz Eq. 1.28.2, set $k = 2\pi/\mathrm{Mpc}$, $A(0) = 1$ and $\dot{A}(0) = 0$. Note that this means we have a wave of wavelength 1 Mpc that starts off at rest with unit amplitude. One can show that in this case $\alpha = 1, \beta = 0$, and the solution for Ψ is therefore

$$\Psi(x,t) = \cos(kvt)\cos(kx) \tag{1.28.5}$$

where $k = 2\pi/\mathrm{Mpc}$. This solution is graphed in the following animation:



Exercise 1.28.2

Check that Eq. 1.28.5 satisfies the wave equation and is consistent with the given initial conditions.

Answer

TBD

We've seen a specific solution to the wave equation. We're now going to work our way toward a completely general solution, and a nifty solution method that relies on the properties we've just seen above for how a cosine (or sine, as we'll see) spatial pattern evolves over time.

Our general solution method will exploit the fact that any some of two solutions to the wave equation is itself a solution to the wave equation.

Exercise 1.28.3

Show that if $\Psi_1(x,t)$ and $\Psi_2(x,t)$ are both solutions of Eq. 1.28.1 then $\Psi_1(x,t) + \Psi_2(x,t)$ is a solution also.

Answer

TBD

Let's now introduce another particular solution to the wave equation, which we will need for the general solutions, and that is:

$$\Psi(x,t) = B(t)\sin(kx). \tag{1.28.6}$$

Exercise 1.28.4

Show that Eq. 1.28.6 is indeed a solution of Eq. 1.28.1 as long as $\ddot{B} = -k^2v^2B$.

Answer

TBD

We are now ready to present the broad outlines of a solution strategy that takes advantage of the fact that any function of x can be written as a sum over cosines and sines of various wavelengths (an assertion that we will discuss more below). The basic idea is that the amplitudes of these sines and cosines will obey a HO equation, and so their time evolution is simple. The general solution is thus a sum over cosines and sines, each with their individual amplitude evolving harmonically at its particular rate.

To be more explicit, here, qualitatively, are the steps:

1) We can write any $\Psi(x,t)$ as a sum over cosines and sines with different wavelengths (and hence different values of k):

$$\Psi(x,t) = A_1(t)\cos(k_1x) + B_1(t)\sin(k_1x) + A_2(t)\cos(k_2x) + B_2(t)\sin(k_2x) + \ldots. \tag{1.28.7}$$

2) If $\Psi(x,t)$ obeys the wave equation then each of the time-dependent amplitudes obeys their own harmonic oscillator equation

$$\ddot{A}_n(t) = -k_n^2 v^2 A(t) \quad \text{and} \quad \ddot{B}_n(t) = -k_n^2 v^2 B(t). \tag{1.28.8}$$

3) These equations for the amplitudes are easy to solve , and their solutions are completely independent of one another: how $A_3(t)$ evolves has no impact on how $B_2(t)$ evolves, for example. (Note that this is kind of amazing because they are both waves in the same medium at the same time and location.)

4) With the time evolution of the amplitudes determined (using the given initial conditions), we can just plug those into Eq. 1.28.7 to get the solution.

One thing we have not told you yet is how one, in practice, actually writes out the terms on the right-hand side of Eq. 1.28.7. For example, how does one know what values of k_1 are needed? Also, how does one get the needed initial conditions for the A_n and B_n? We'll get to that, but for now let's look at an example of this solution method at work.

Let's work out how $\Psi(x,t)$ will evolve if it starts off as a triangle wave at rest. Let's assume the triangle wave has a wavelength of 1 Mpc, initially has an amplitude of unity, is initially at rest ($\dot{\Psi}(x,0) = 0$) and is phased so that it is zero at the origin ($\Psi(0,0) = 0$). Let's further assume it obeys the wave equation with speed v; i.e. Eq. 1.28.1.

We will state without proof here (but the proof is not difficult; see Wolfram Alpha) that the initial configuration $\Psi(x,0)$ can be written as

$$\Psi(x,0) = \Sigma_{n=1}^{\infty} \left(A_n(0)\cos(k_n x) + B_n(0)\sin(k_n x)\right) \tag{1.28.9}$$

with $A_n(0) = 0$ for all n, $B_n(0) = 0$ for even n, and $B_n(0) = 8/\pi^2(-1)^{(n-1)/2}/n^2$ for odd n and their time derivatives at $t = 0$ vanishing. In the figure we show how well the triangle wave is approximated by the series as we increase the number of terms we are including in the sum, by increasing the maximum value of n, so you can see that this series representation does indeed seem to work. In addition to the sum, we also show the individual terms.

Lucas, please insert a still (non-animated) figure here showing the series converging

To be able to explicitly show the solutions, and so this is not too cumbersome, we will restrict ourselves from here on out to just the first three terms in the sum. From the initial conditions written above we thus have

$$\begin{aligned}\Psi(x,t) = \frac{8}{\pi^2}\cos\left(\frac{2\pi}{Mpc}vt\right)\sin\left(\frac{2\pi}{Mpc}x\right) - \frac{8}{9\pi^2}\cos\left(\frac{6\pi}{Mpc}vt\right)\sin\left(\frac{6\pi}{Mpc}x\right) \\ + \frac{8}{25\pi^2}\cos\left(\frac{10\pi}{Mpc}vt\right)\sin\left(\frac{10\pi}{Mpc}x\right) + ...\end{aligned} \tag{1.28.10}$$

The solution is illustrated in the animation.

Lucas, please insert an animated figure here as described immediately above.

**Lloyd: insert here some wrap-up of above section **

The Continuous Fourier Transform

We just saw a solution for an initial spatial configuration with wavelength $\lambda = 1$ Mpc which can be represented as a sum over sines and cosines (just sines in this case) with an (infinite) set of discrete k values, specifically $k_n = 2n\pi/\lambda$. Note that the spacing between k values in this case is $\Delta k = 2\pi/\lambda$. For the more general situation of a function of space that is not periodic, we can think of it is a periodic function with infinite wavelength. As the wavelength goes to infinity, the Δk goes to zero. So we see we need a continuum of values of k. For the general case then we swap the sum over i with an integral over k:

$$\Psi(x,t) = \int_0^{\infty} dk[A(k,t)\cos(kx) + B(k,t)\sin(kx)]. \tag{1.28.11}$$

It turns out there is a more compact way of working with this decomposition into cosines and sines if we use complex numbers. We can write instead

$$\Psi(x,t) = \int_{-\infty} dk \Psi(k,t) e^{ikx} \tag{1.28.12}$$

which is a mathematical opertation known as the inverse Fourier transform. For $\Psi(x,t)$ a real function, Eq. 1.28.11 and Eq. 1.28.12 are equivalent if we make the identification

$$A(k,t) = 2Re\tilde{\Psi}(k,t) \quad \text{and} \quad B(k,t) = 2Im\Psi(k,t) \tag{1.28.13}$$

for $k > 0$ where "Re" and "Im" indicate taking the real and imaginary parts respectively. Homework problem TBD is to prove these relationships are true.

Solving the Wave Equation in Fourier Space

You may already be familiar with a method for solving partial differential equations known as separation of variables. Using separation of variables to solve the wave equation, we would guess a solution of the form $\Psi(x,t) = X(x)T(t)$. Plugging this into the wave equation yields two simple ODE's: one for $T(t)$ and one for $X(x)$. Now though, we'd like to introduce you to another way to analyze partial differential equations (PDE's): Fourier methods.

The basic idea here is that we transform from a basis in which the time evolution is complicated (one in which the field is described as a function of position), to a basis in which the time evolution is remarkably simple (one in which the field is described as a collection of Fourier modes). We do the time evolution in this new basis, and then we transform back to our original basis.

We will use the discrete version of the Fourier transform here, as that is perhaps an easier starting point to wrap one's mind around first. We include a discussion of the continuous Fourier transform, which is easy to understand as the continuum limit of the discrete version.

[To be done: all this needs to be translated to discrete from continuous and then we need to create a section on the continuum limit.]

We start off, in a manner that may seem a little backwards, by defining the inverse Fourier transformation:

$$h(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk e^{ikx} \tilde{h}(k). \tag{1.28.14}$$

The $\tilde{h}(k)$ are complex (have real and imaginary parts) and recall that $\exp(ikx) = \cos(kx) + i\sin(kx)$. We start here because there is a theorem that states that a broad class of functions of x can all be written as sums over $\exp(ikx)$ for a continuum of values of k, and for appropriately chosen complex coefficients of the $\exp(ikx)$. That is, we can represent the information in a function $h(x)$ by its Fourier coefficients $\tilde{h}(k)$, with the relationship between the two given by Equation1.28.14. The functions, $\exp(ikx)$ are known as Fourier modes. Since $\exp(ikx) = \exp(ik(x + 2\pi/k))$ we see that a Fourier mode has a wavelength of $2\pi/k$. We call k the 'wavenumber.'

One can do Fourier transforms in time or in space or both. Here we are only going to be doing Fourier transforms in space, although we will consider Fourier transforms in space *at all points in time*. To be explicit about this, we can rewrite Equation 1.28.14 to include a t argument of the functions:

$$h(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk e^{ikx} \tilde{h}(k,t). \tag{1.28.15}$$

It's the same transformation, but now we are explicit that we do this transformation at all values of t.

Recall that we claimed that the evolution of the $\tilde{h}(k,t)$ would be simple. To figure out what equation governs the evolution of these coefficients, we need to know how to figure out for a given $h(x,t)$ what is $\tilde{h}(k,t)$. But we are going to return to leaving off the t dependence, for simplicity. We already know how to go from $\tilde{h}(k)$ to $h(x)$, that is what we called the inverse Fourier transform, Equation 1.28.14. So we are looking now for the inverse of this, what we will naturally call the Fourier transform.

Let's work our way toward the Fourier transform by first pointing out an important property of Fourier modes: they are *orthonormal*. This means that if we integrate over all space one Fourier mode, e^{-ikx}, multiplied by the complex conjugate of another Fourier mode $e^{ik'x}$ the result is 2π times the Dirac delta function:

$$\int_{-\infty}^{\infty} dx e^{-ikx} e^{ik'x} = 2\pi\delta(k - k') \tag{1.28.16}$$

where the Dirac delta function is a continuum version of the Kronecker delta function, defined by its integral over k such that

$$\int_{-\infty}^{\infty} dk\delta(k-k')f(k) = f(k'). \tag{1.28.17}$$

You can loosely think of the Dirac delta function as being zero for all non-zero values of its argument and $+\infty$ when its argument is zero.

From these equations one can derive what we call the Fourier transform:

$$\tilde{h}(k) = \int_{-\infty}^{\infty} dx e^{-ikx} h(x) \tag{1.28.18}$$

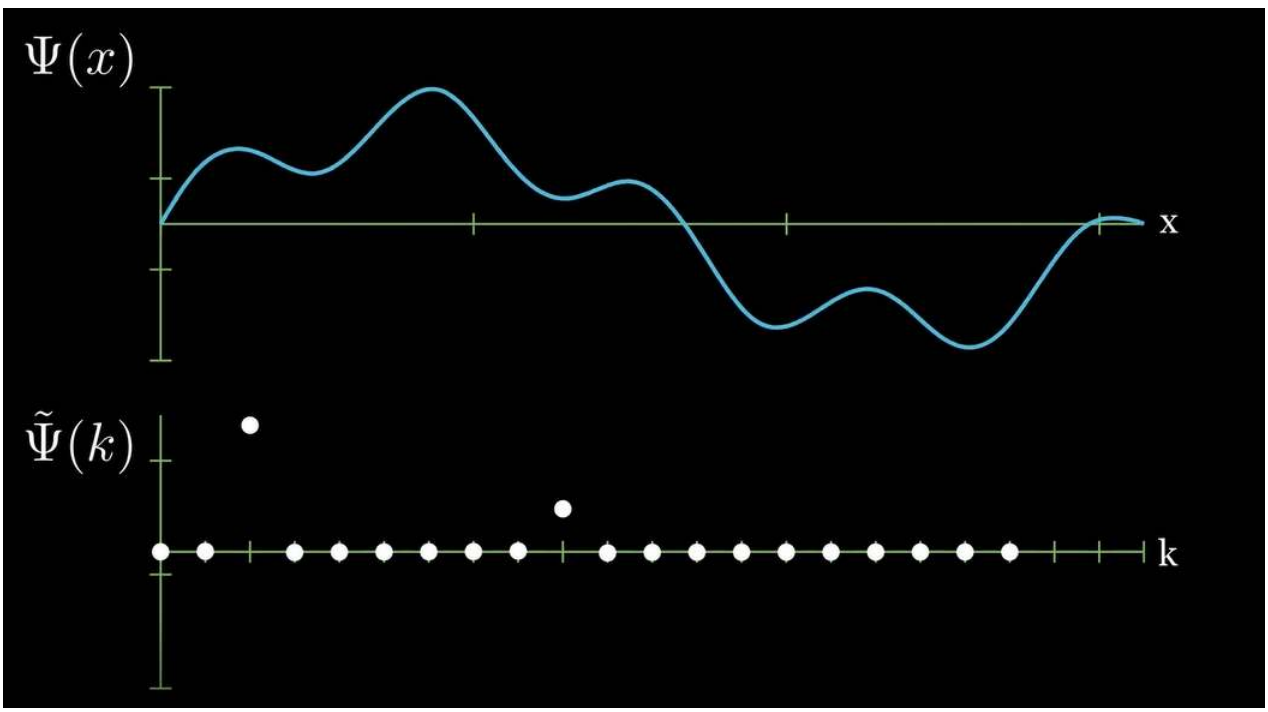
and thus the answer to the question of how we deduced $\tilde{h}(k,t)$ from $h(x,t)$.

Box 1.28.6

Exercise 27.1.1: Show that one can derive Equation 1.28.18 from Equations 1.28.14, 1.28.16, and 1.28.17.

Before deriving the evolution equation for the Fourier coefficients, let's look at an example of a function in the position basis and what it looks like in the Fourier basis. The following image shows a wave on the top panel, $\Psi(x)$, and the Fourier transform of that wave on the bottom panel. (Note that $\mathcal{F}(\Psi)$ indicates the operation of Fourier transforming the function $\Psi(x)$; i.e., $\mathcal{F}(\Psi) = \tilde{\Psi}(k)$. Notice how the Fourier transform 'picks out' the two spatial frequencies of which the wave is composed.

[Problem: this is a discreet FT and we have only talked about continuum.]



For a $\Psi(x,t)$ that obeys the wave equation, let's now find the equation that its Fourier coefficients, $\tilde{\Psi}(k,t)$, satisfy. Starting from the wave equation,

$$\frac{\partial^2 \Psi(x,t)}{\partial t^2} = v^2 \frac{\partial^2 \Psi(x,t)}{\partial x^2}, \tag{1.28.19}$$

and then substituting in the inverse Fourier transform $\Psi(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dk\tilde{\Psi}(k,t)e^{-ikx}$ we find:

$$\frac{\partial^2}{\partial t^2}\int_{-\infty}^{\infty} dk\tilde{\Psi}(k,t)e^{-ikx} = v^2\frac{\partial^2}{\partial x^2}\int_{-\infty}^{\infty} dk\tilde{\Psi}(k,t)e^{-ikx} \tag{1.28.20}$$

Distributing the derivatives gives:

$$\int_{-\infty}^{\infty} dk\frac{\partial^2\tilde{\Psi}(k,t)e^{-ikx}}{\partial t^2} = -\int_{-\infty}^{\infty} dk(kv)^2\tilde{\Psi}(k,t)e^{-ikx}. \tag{1.28.21}$$

We can then rearrange terms to find:

$$\int_{-\infty}^{\infty} dk\left[\frac{\partial^2\tilde{\Psi}(k,t)}{\partial t^2} + (kv)^2\tilde{\Psi}(k,t)\right]e^{-ikx} = 0. \tag{1.28.22}$$

It turns out that the only way the left-hand side can be zero for all values of x is if the quantity in square brackets is zero for all values of k (see Box below) so we get that

$$\frac{\partial^2\tilde{\Psi}(k,t)}{\partial t^2} + (kv)^2\tilde{\Psi}(k,t) = 0. \tag{1.28.23}$$

Box 1.28.7

Exercise 27.2.1: Prove that if

$$\int_{-\infty}^{\infty} dkf(k)e^{-ikx} = 0 \tag{1.28.24}$$

for all x, then $f(k) = 0$ for all k.

First, multiply the left-hand side of Equation 1.28.24 by $\exp(-ik'x)$, integrate it over all k', and identify the Dirac delta function to end up with:

$$\frac{\partial^2\tilde{\Psi}(k')}{\partial t^2} + (k'v)^2\tilde{\Psi}(k') = 0. \tag{1.28.25}$$

Finally, note that since this is true for all k' it's also true for all k.

Equation 1.28.23 is a very common differential equation. You've probably solved it many times! You may recognize it better if we let $y = \tilde{\Psi}(k,t)$, so that it reads $\ddot{y} + k^2v^2y = 0$. We can easily write down a solution:

$$\tilde{\Psi}(k,t) = A(k)\sin(kvt) + B(k)\cos(kvt). \tag{1.28.26}$$

Thus our general solution back in the space basis is

$$\Psi(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dk\Big[A(k)\sin(kvt) + B(k)\cos(kvt)\Big]e^{ikx}. \tag{1.28.27}$$

We can find $A(k)$ and $B(k)$ if we know $\Psi(x,t)$ and $\dot{\Psi}(x,t)$ at $t = 0$ because

$$\Psi(x,t=0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dkB(k)e^{ikx} \tag{1.28.28}$$

and

$$\dot{\Psi}(x,t=0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dkkvA(k)e^{ikx} \tag{1.28.29}$$

.

Given these relationships we see that to get $B(k)$ and $A(k)$ we Fourier transform the initial value of Ψ and its time derivative:

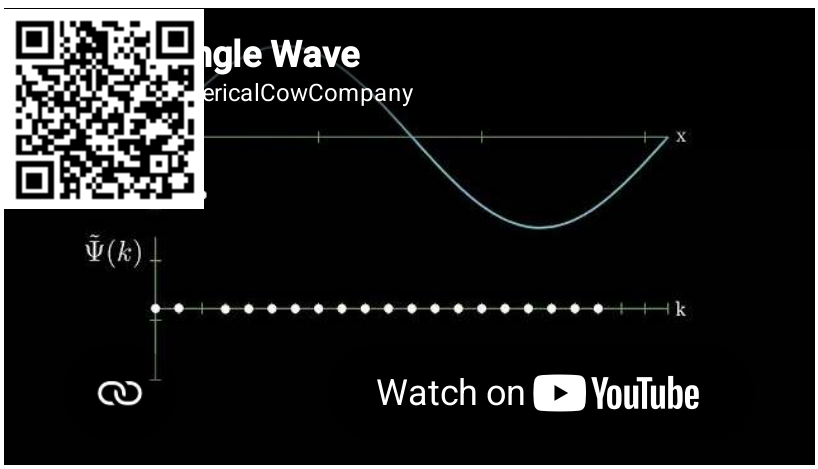
$$B(k) = \int_{-\infty}^{\infty} dx \Psi(x, t = 0) e^{-ikx} \tag{1.28.30}$$

and

$$A(k) = \frac{1}{kv} \int_{-\infty}^{\infty} dx \dot{\Psi}(x, t = 0) e^{-ikx}. \tag{1.28.31}$$

To summarize, we found that in a Fourier basis, rather than the original space basis, the wave equation simplifies from a partial differential equation to a set of uncoupled ordinary differential equations. The wave equation is easily solved in the Fourier basis and we provided the general solution. This general solution depends on two functions of k that can be derived from the initial conditions.

Consider the following initial conditions on our string $\Psi(x, t = 0) = \sin(2x)$. This is a single wave with k = 2. Taking the Fourier transform, we find: $\mathcal{F}\Big(\Psi(x, t = 0)\Big) = \delta(x - 2)$. The Fourier transform is 1 where k = 2 and 0 otherwise. We see that over time, the amplitude of this wave oscillates with cos(2 v t). The solution to the wave equation for these initial conditions is therefore $\Psi(x, t) = \sin(2x)\cos(2vt)$. This wave and its Fourier transform are shown below. The power spectrum is merely the Fourier transform squared.



Now consider we have initial conditions which are more complicated, but can be written as an infinite sum of sine waves as follows:

$$\Psi(x, t = 0) = \sum_{i=1}^{\infty} A_i \sin(k_i x) \tag{1.28.32}$$

Taking the Fourier transform, we find the following sum of delta functions:

$$\mathcal{F}\Big(\Psi(x, t = 0)\Big) = \sum_{i=1}^{\infty} A_i \delta(k - k_i) \tag{1.28.33}$$

Which oscillate in time according to:

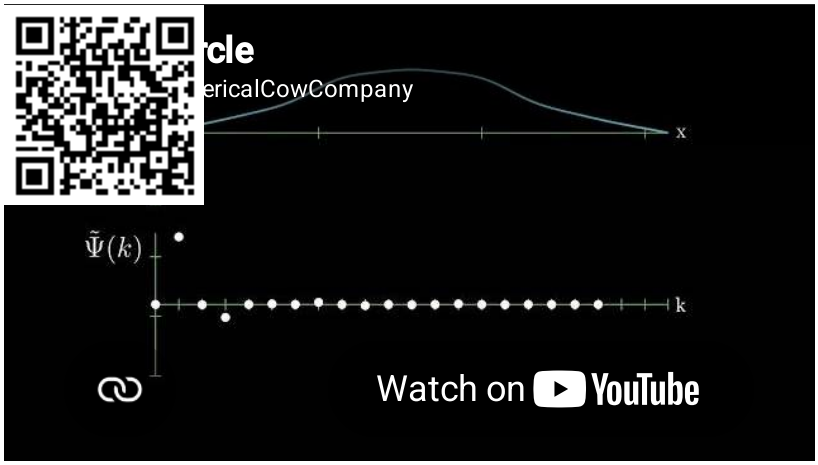
$$\mathcal{F}\Big(\Psi(x, t)\Big) = \sum_{i=1}^{\infty} A_i \delta(k - k_i) \cos(k_i vt) \tag{1.28.34}$$

Returning to real space we find:

$$\Psi(x, t) = \sum_{i=1}^{\infty} A_i \sin(k_i x) \cos(k_i vt) \tag{1.28.35}$$

The takeaway here is that the solution to the wave equation can always be written as a sum of independent standing waves. Some examples are shown below. The top panel shows the wave and the bottom panel shows the Fourier transform of that wave. Notice

how the evolution seems very complex in real space, but in Fourier space it is merely independent delta functions of oscillating amplitude. This is the beauty of using Fourier methods to analyze the wave equation. If you wanted to see the power spectrum, you would simply square the Fourier transform.



Exercise 1.28.8

Consider the heat equation for a straight rod: $\frac{d\Psi}{dt} = \alpha \frac{d^2\Psi}{dt^2}$, where $\Psi(x,t)$ is the temperature at a certain point on the beam. Using the techniques from the previous section, find the evolution of Fourier modes. How can this physically be interpreted?

Answer

We plug in the Fourier representation of Ψ into the heat equation:

$\frac{d}{dt}\int_{-\infty}^{\infty} \mathcal{F}(\Psi)e^{-ikx}dk = \alpha \frac{\partial^2}{\partial x^2}\int_{-\infty}^{\infty} \mathcal{F}(\Psi)e^{-ikx}dk$

Distributing the derivatives and some algebra gives:

$\int_{-\infty}^{\infty}\left[\frac{d\mathcal{F}(\Psi)}{dt}e^{-ikx} + \alpha k^2 e^{-ikx}\mathcal{F}(\Psi)\right]dk = 0$

Which is satisfied if:

$\frac{d\mathcal{F}(\Psi)}{dt} + \alpha k^2 \mathcal{F}(\Psi) = 0$

Using separation of variables, we find:

$\mathcal{F}(\Psi) = Ce^{-\alpha k^2 t}$, where C is a constant determined by initial conditions.

Therefore, we see that higher frequencies decay faster. This makes sense, as we would expect spikes in temperature (high curvature) to disappear quickly, whereas more smooth temperature gradients will decay more slowly.

This page titled 1.28: Solving the Wave Equation with Fourier Transforms is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.29: The First Few Hundred Thousand Years- The Dynamics of the Primordial Plasma

Introduction

This chapter is Under Construction You probably noticed in the previous chapter that the power spectrum of the CMB has a series of peaks in it. In this chapter we will explain the origin of those peaks as arising from the acoustic dynamics in the primordial plasma.

Waves in the Plasma

The microphysical composition of the plasma is not too important for our presentation here. What is important is that we can model it as a fluid. We can model a system of particles as a fluid when they rapidly scatter off of each other. The validity of the fluid approximation is a matter of length scale, becoming valid on scales that are large compared to the mean free path, the typical distance traveled by a particle before being scattered by another particle. On sufficiently large scales we can ignore the details of the trajectories of individual particles and model the system as being completely defined everywhere by a density, pressure, and velocity at every point in space and time. When we do so, we say we are modeling the medium as a fluid. Another example of a system of particles that can be well-approximated as a fluid is the air in the room you are in, where the mean free path of a nitrogen molecule is about 10^{-5} cm.

The primordial plasma was extremely uniform, with density varying from one place to another by as little as about 0.001%. That's what we meant earlier by 'gently disturbed away from equilibrium.' Very gently! Associated with these small variations in density are both variations in the pressure and gravitational potential. Gradients in the pressure and gravitational potential result in forces on the plasma that drive the evolution of its density and velocity. If we ignore gravity (and the expansion of space), the dynamics of the plasma are governed by a simple wave equation:

$$\frac{\partial^2 \Psi}{\partial t^2} = c_s^2 \nabla^2 \Psi \tag{1.29.1}$$

where $\Psi = \delta\rho$ is the plasma density minus the spatially averaged plasma density, $c_s^2 = \partial P/\partial\rho$ is the square of the sound speed in the plasma, and P is the pressure of the plasma. Recall also that $\nabla^2\Psi = \frac{\partial^2\Psi}{\partial x^2} + \frac{\partial^2\Psi}{\partial y^2} + \frac{\partial^2\Psi}{\partial z^2}$.

Although gravity and expansion do play very important roles in the behavior of the plasma, we postpone the discussion of these complications until later. For now, this simple equation is sufficient for a qualitative understanding of the dynamics of the plasma and the origin of the acoustic peaks.

To give you some feel for how the density evolves under this wave equation, we show here how a localized spherical overdensity evolves in time, assuming the fluid initially has zero velocity. The pressure gradients drive the fluid outward in a shell -- seen here in this 2-dimensional slice as a ring.

Acoustic Oscillations in a Guitar String

Before further discussion of the ancient plasma that filled the infant universe, let's consider a system that's a bit closer to home: guitar strings. Although these two systems may seem very different, they have similar dynamics. Both are governed by the same wave equation.

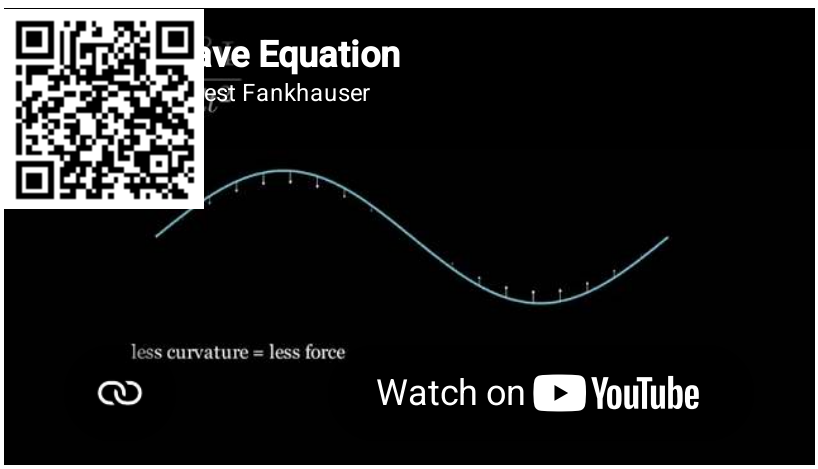
Introducing the Wave Equation

For the guitar string, in just one dimension:

$$\frac{\partial^2 \Psi(x,t)}{\partial t^2} = v^2 \frac{\partial^2 \Psi(x,t)}{\partial x^2} \tag{1.29.2}$$

Here, $\Psi(x,t)$ is the displacement of the guitar string at a given point along the string and v is the velocity of a wave traveling on the guitar string (determined by the tension and density of the string).

We won't take the time to derive the wave equation, but instead we'd like to give some intuition for where it comes from. Consider the segment at the center of the guitar string, $x = \frac{1}{2}L$, where L is the length of the string. Then, $\frac{\partial^2 \Psi(L/2,t)}{\partial t^2}$ is the acceleration of that segment, which is proportional to the force on that segment. Recall that the second derivative with respect to space, $\frac{\partial^2 \Psi(x,t)}{\partial x^2}$ is related to the concavity of the segment. That is, if $\frac{\partial^2 \Psi(L/2,t)}{\partial x^2}$ is large, the string is very bent at the center. If $\frac{\partial^2 \Psi(L/2,t)}{\partial x^2}$ is zero, then the string is straight. The wave equation states that the force on a segment of string is proportional to the curvature of the string at that point. To make this more clear, watch the following animation. Does the idea that force is proportional to curvature match your intuition?



There are two important takeaways from the video. First, we see that the force on the string at a given x is proportional to the curvature at that point. This makes intuitive sense! Second, we notice that since the force is higher, strings with higher curvatures oscillate faster. Also notice that the curves with high curvature have smaller features. As the width of the bump shrinks, the curvature and force increase. Likewise, the sine wave with a smaller wavelength has greater curvatures and forces. This is important to understanding the power spectrum of the CMB, as features of the CMB with smaller angular scale oscillate faster.

Acoustic Oscillations in the Primordial Plasma

Here we explain the existence of the bumps and wiggles in the power spectrum.

Simplified Evolution Equation

The photon background at any point in the sky has some temperature; we call this the temperature monopole, the $\ell = 0$ mode, and denote its fractional departure from the mean temperature as $\Theta_0 \equiv (T - \bar{T})/\bar{T}$ where the bar here indicates an average over all space. Note that Θ_0 is a function of time and space and the zero subscript refers to the monopole aspect; it is not indicating the current epoch. Here we are going to write down a simplified equation for the evolution of this monopole field under the influence of pressure gradients.

Since the photons have a black body distribution, there is a relationship between temperature and density, $\rho_\gamma \propto T^4$, which leads to $\delta\rho_\gamma/\rho_\gamma = 4\Theta_0$. The main thing to note is that the the density and temperature of the plasma are proportional. More dense plasma is hotter, whereas less dense regions will be cooler. So as we are evolving the temperature Θ_0 we are also evolving the plasma density.

Like the guitar string, the temperature perturbations obey the wave equation. Moving to Fourier space and using the notation $\tilde{\Theta}_0 = \mathcal{F}(\Theta_0)$, we have the same equation we found previously:

$$\frac{d^2\tilde{\Theta}_0}{dt^2} = -k^2 c_s^2 \tilde{\Theta}_0 \tag{1.29.3}$$

where c_s is the speed of sound in the plasma. The speed of sound in the plasma is related to properties of the medium it travels through: $c_s^2 \equiv \partial P/\partial \rho$.

So in Fourier space, the plasma dynamics are given by:

$$\tilde{\Theta}_0 = \mathcal{F}\Big(\Psi(x, t=0)\Big)\cos\left(kc_s t\right) \tag{1.29.4}$$

Moving from One to Three Dimensions

The guitar string was a great starting point, because it was a one-dimensional case (the string's displacement is a function only of x). On the other hand, the primordial plasma filled the entire universe, so Θ_0 is a function of x, y, and z. Luckily, the same tools we used before easily extend to more dimensions.

Now, let $\vec{x} = \langle x, y, z\rangle$ and $\vec{k} = \langle k_x, k_y, k_z\rangle$.

The Fourier transform is now defined as:

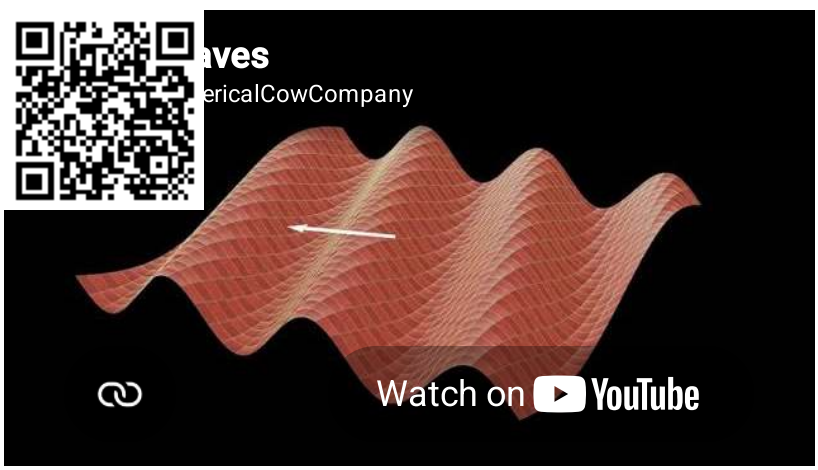
$$\mathcal{F}(h(\vec{x})) = g(\vec{k}) = \int_{-\infty}^{\infty} e^{i\vec{k}\cdot\vec{x}} h(\vec{x}) dxdydz \tag{1.29.5}$$

and the inverse Fourier transform likewise:

$$\mathcal{F}^{-1}(g(\vec{k})) = h(\vec{x}) = \int_{-\infty}^{\infty} e^{-i\vec{k}\cdot\vec{x}} g(\vec{k}) dk_x dk_y dk_z \tag{1.29.6}$$

Notice that to extend to more dimensions, we simply use vectors and dot products instead of scalars and multiplication.

You might be a little uncertain what it means for k, the 'wave number', to be a vector. In multiple dimensions, the magnitude of $\vec{k}$ determines the wavelength and the direction of $\vec{k}$ is the direction of propagation of the wave. The example below shows the wave $\sin(\vec{k}\cdot\vec{x})$ (in red checkerboard) for many different values of $\vec{k}$ (the white arrow).



Initial Conditions of the Plasma

The last key to understanding the evolution of the primordial plasma is the initial conditions. The initial power spectrum was random noise with a power spectrum of $1/k^3$. The random noise came from quantum fluctuations that were amplified and 'baked-in' to the plasma during inflation. The power spectrum was initially $1/k^3$ because the plasma was scale invariant. A scale invariant plasma has the following property: imagine taking a snapshot of a 1 Mpc by 1 Mpc area of plasma and another snapshot of a 1,000 Mpc by 1,000 Mpc area; the two snapshots would be indistinguishable.

Using this information, we can now make artificial initial conditions for the primordial plasma in 3-dimensions. The procedure is as follows:

1. Create a 3D grid of white noise. These are the random fluctuations in the plasma.
2. Fourier transform the white noise to move to Fourier space.
3. Multiply the grid by $1/k^{3/2}$ to get the right power spectrum. Note that $k = \|\vec{k}\| = \sqrt{k_x^2 + k_y^2 + k_z^2}$
4. Inverse Fourier transform the grid to move back to real space. We now have some artificial initial conditions for the plasma density.

A Python script that does this procedure in two dimensions is shown below. Try running it a few times to see several different initial conditions. Notice that while they are each unique, they all have the same statistical properties.

```
import numpy as np
import matplotlib.pyplot as plt

resolution = 2**8 # Number of pixels on each side of map

# Creates a 2D grid of noise in Fourier space
noise = np.random.normal(0,1, size = (resolution,resolution))
noise_ft = np.fft.fft2(noise)

# Defines K-vector over the grid
KX, KY = np.meshgrid(np.linspace(-1, 1, resolution), np.linspace(-1, 1, resolution))
K_magnitude = np.sqrt(KX**2 + KY**2)

# Give the noise the power spectrum we want
fourier_space_grid = K_magnitude**(-3/2) * noise_ft

# Use the inverse Fourier transform to create a CMB map in real space
CMB_map = np.fft.ifftn( np.fft.fftshift(fourier_space_grid)).real

# Plots and displays CMB map
plt.axis('off')
plt.imshow(CMB_map, cmap = 'plasma')
```

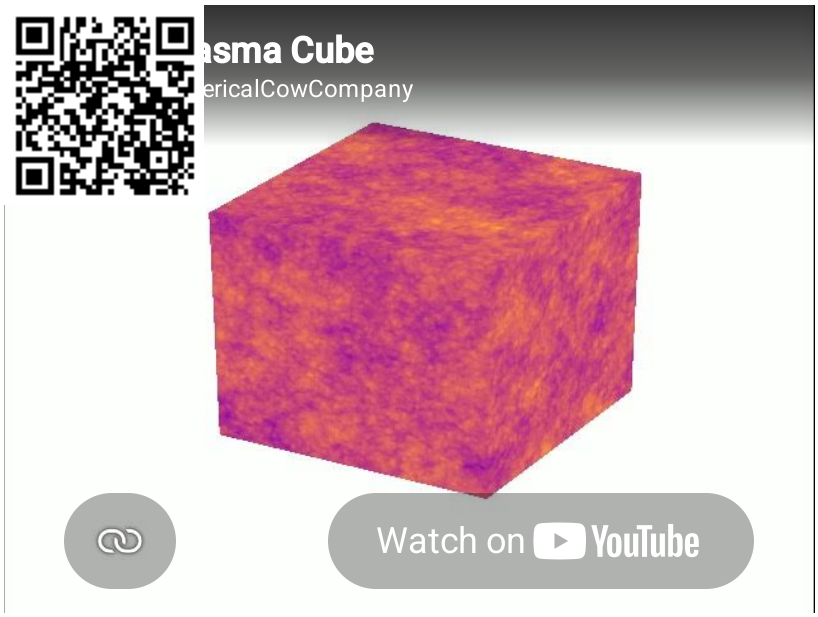
run | run all | restart | restart & run all

```
<matplotlib.image.AxesImage at 0x7faffbd78690>
```

Evolving the Plasma

We have both initial conditions and a solution to the wave equation. Our last step is to tie everything together so we can see the plasma evolve. This is actually quite simple. Using the same technique as the previous section, we generate initial conditions. Then, in Fourier space we multiply our initial conditions by $\cos(\|\vec{k}\| c_s t)$ to get $\tilde{\Theta}_0$. Finally, we inverse Fourier transform $\tilde{\Theta}_0$ so that we have a solution in real space, Θ_0.

Using this technique, we've created an animation of the primordial plasma evolving below. In order to visualize the opaque plasma, we've removed all but a cube of plasma with side lengths of 1024 Mpc. It is amazing that the simple wave equation can lead to such complexity and beauty.



Evolution of the Power Spectrum

Below, we show an animation of the same cube of plasma, but now the power spectrum is also shown on the right side. We can see that lower frequencies (on the left side of the x-axis) oscillate more slowly, whereas higher frequencies (on the right side) oscillate more quickly. Over time, this creates the peaks and dips that we see in the power spectrum. (Note, the video loops several times)

In this chapter, we've come all the way from the simple wave equation to an understanding of the mechanism which created the peaks and troughs in the power spectrum of the cosmic microwave background. Also, we know how to use Fourier methods to solve linear partial differential equations.

This page titled 1.29: The First Few Hundred Thousand Years- The Dynamics of the Primordial Plasma is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.30: Structure Formation

Shortly before the epoch of the CMB, matter became the dominant constituent of the universe. Throughout the matter-dominated epoch---i.e., from a few hundred thousand years after the Big Bang until dark energy--matter equality nearly 13 billion years later---dark matter and baryonic structure formed and grew. This Chapter describes how, under the gravitational influence of dark matter, small fluctuations in the matter density field evolve. Particularly dense regions collapse into nonlinear, self-gravitating systems called *dark matter halos*, which form the nodes of the cosmic that cosmologists observe today.

Linear Structure Formation

A key quantity used to describe dark matter structure is the *density field* $\rho(\vec{r}, a)$, which specifies the matter density ρ (i.e., the mass contained in some volume V) as a function of three-dimensional position $\vec{r}$ and scale factor a. Specifically, consider the *density contrast*

$$\delta(\vec{r}, a) = \frac{\rho(\vec{r}, a) - \bar{\rho}(a)}{\bar{\rho}(a)}, \tag{1.30.1}$$

where $\bar{\rho}(a)$ is the mean matter density at a given epoch. The density contrast describes how matter is distributed relative to the mean density; at a given location and time, $\delta > 0$ ($\delta < 0$) corresponds to an overdense (underdense) region, and $\delta = 0$ corresponds to a region of average density. We refer to structure on a given scale as *linear* when $|\delta| \ll 1$ and *nonlinear* when $|\delta| \approx 1$. Linear structure formation can largely be modeled analytically, while nonlinear structure formation requires numerical simulations to model accurately.

How does δ evolve over time, and what are its typical values at early times and today? Recall the Friedmann equation from Chapter 11:

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\bar{\rho}}{3}, \tag{1.30.2}$$

where we assume that the universe is flat ($k = 0$) and use $\bar{\rho}$ to denote the density is averaged over a large region of the universe. Technically, the density that enters the Friedmann equation is the sum of matter, radiation, and dark energy, but this sum is dominated by matter during the matter-dominated epoch.

Consider a region of the universe with some local matter density ρ. That region will either be overdense and behave locally like a closed universe, or underdense and behave locally like an open universe. Thus, this region obeys

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\rho}{3} - \frac{k}{a^2} \tag{1.30.3}$$

$$\implies \rho - \bar{\rho} \propto \frac{k}{a^2} \tag{1.30.4}$$

$$\implies \delta = \frac{\rho - \bar{\rho}}{\bar{\rho}} \propto \frac{1}{\bar{\rho}a^2} \propto \frac{a^3}{a^2} \propto a, \tag{1.30.5}$$

where Equation 1.30.5 follows by subtracting Equations 1.30.4 and 1.30.3, and the final line uses $\bar{\rho} \propto a^{-3}$ as appropriate for non-relativistic matter. For a more rigorous derivation of this scaling, see Problem 31.1.

As the universe expands, overdense regions with $\delta > 0$ become denser relative to the background according to Equation 1.30.5. Gravity attracts matter surrounding a given overdensity towards its center, slowing down the expansion of overdense regions relative to the background and causing their density contrast to grow.

Box 1.30.1

Exercise 31.1.1: Qualitatively describe how underdense regions evolve in the linear regime. Does their density contrast increase or decrease as the universe expands? Justify your answer mathematically.

At the time of the CMB, baryon density fluctuations have an average size of

$$\delta_{\text{baryon}}(a_{\text{CMB}}) \sim \frac{\Delta\rho_{\text{baryon}}}{\bar{\rho}_{\text{baryon}}} \sim \frac{\Delta T_{\text{baryon}}}{\bar{T}_{\text{baryon}}} \approx 10^{-5}. \tag{1.30.6}$$

According to Equation 1.30.5, the average size of these fluctuations today is therefore

$$\delta_{\text{baryon}}(a_{\text{today}}) \approx \frac{a_{\text{today}}}{a_{\text{CMB}}}\delta_{\text{baryon}}(a_{\text{CMB}}) \approx 10^3\delta_{\text{baryon}}(a_{\text{CMB}}) \approx 10^{-2}. \tag{1.30.7}$$

This is **not** what we observe! Instead, today's universe is filled with galaxies, which have densities larger than the background by orders of magnitude. In other words, large, non-baryonic density fluctuations must be present at the time of the CMB in order for nonlinear structure to form by today. These density fluctuations are sourced by dark matter. Problem 31.2 explores why baryonic perturbations are much smaller than dark matter perturbations at the time of the CMB.

Dark matter drives the growth of small baryonic overdensities observed in the CMB (left) into nonlinear structure traced by galaxies today (right).

Nonlinear Structure Formation

What happens to overdensities as they near $\delta \approx 1$? Consider a toy model in which a region of the universe has a uniformly higher matter density than the rest,

$$\rho(r, t_i) = \bar{\rho}(t_i)\begin{cases}1+\delta_i, r < R_i \\ 1, r > R_i\end{cases} \tag{1.30.8}$$

where r is the radial distance from the center of the overdensity, R_i is the initial size of the overdensity, t_i is the initial time, $\bar{\rho}(t_i)$ is the background density, and δ_i is the size of the overdensity. This is referred to as a *top-hat overdensity.*

Problem 31.3 explores the evolution of this system. The overdensity initially expands with the Hubble flow; if $\delta_i > \Omega_m(t_i)^{-1} - 1$, gravitational attraction overwhelms the expansion, and the region collapses. This picture schematically describes how self-gravitating *dark matter halos* form.

Box 1.30.2

Exercise 31.2.1: Realistic cosmological density fluctuations are more complex than the top-hat model for several reasons:

- Overdensities are not spatially uniform; instead, the density contrast is higher near the centers of overdensities and decreases until joining onto the mean density;
- Fluctuations are not spherically symmetric; instead, collapse occurs successively along the major, intermediate, and minor axes of the initial perturbation;
- Collapse is not purely radial; particles orbiting an overdensity have angular momentum.

Qualitatively describe how each effect influences how overdensities evolve. Assuming that the overdensity will collapse, do you expect each effect to speed up or slow down the process?

Gravitational collapse naturally leads to *bottom-up* (or *hierarchical*) structure formation. Consider a shell of matter at distance r from the center of an overdensity as it collapses after turnaround. The time taken for this shell to collapse is given by

$$t_{\rm ff} \sim \sqrt{r/g} \sim \sqrt{r^3/(GM)} \sim \sqrt{1/(G\rho)}, \tag{1.30.9}$$

where M is the enclosed mass and ρ is the enclosed density. This timescale, referred to as the *free-fall time*, is relevant in many other astrophysical settings where gravity plays an important role.

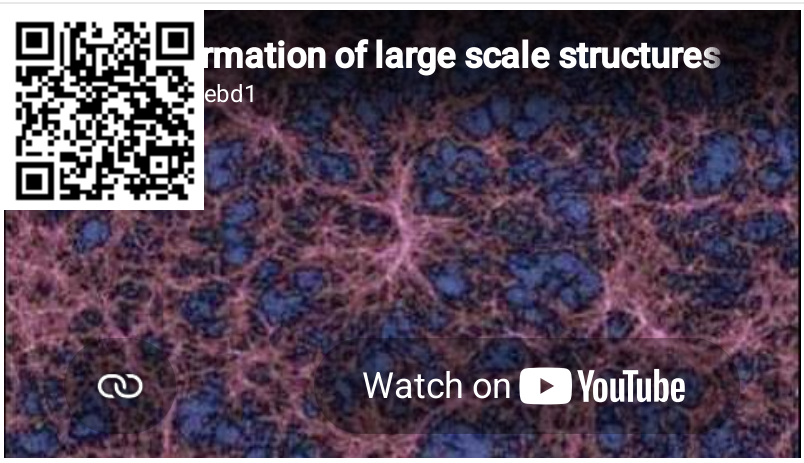
Matter fluctuations have higher *average* densities at early times because the background matter density is higher (even though *overdensities* are smaller at early times). Thus, according to Equation 1.30.9, small overdensities at early times collapse more quickly than large overdensities at later times. These small overdensities collapse, forming small dark matter halos, which merge together to form ever larger structures. The typical masses of these first-generation halos depend on the properties of dark matter and the primordial power spectrum. They are typically at least 10 orders of magnitude smaller than the most massive ($\approx 10^{15}\ M_\odot$) clusters today; in many models, including weakly interacting massive particle (WIMP) dark matter, the first halos are even smaller ($\approx 10^{-6}\ M_\odot$). In contrast, theories of neutrino dark matter considered in the 1970s predict *top-down* structure formation, in which "superclusters" form and subsequently fragment into smaller pieces. This scenario is ruled out by observations, which show that small galaxies and halos form first and subsequently merge to form massive objects.

Numerical simulations are used to predict the distribution and properties of halos over a wide range of mass scales and redshifts. In particular, cosmological *N-body simulations* solve for the gravitational evolution of discretized particles, which represent the dark matter density field, in an expanding FLRW universe. A key prediction of these simulations is the mass distribution of collapsed halos, or the *halo mass function*, which is predicted to increase as halo mass decreases in a roughly scale-invariant fashion:

$$\frac{dn}{dM} \propto M_{\rm halo}^{-\alpha}, \tag{1.30.10}$$

where n is the comoving number density of dark matter halos, M is halo mass, and $\alpha \approx 1.9$.

Of course, it's difficult to compare predictions for the distribution of dark matter halos directly to observations. Chapter 32 explores how galaxies form in dark matter halos, and shows that the hierarchical structure formation paradigm describes observations of the nonlinear galaxy distribution today extremely well.



Growth of dark matter structure in a cosmological N-body simulation. Pink (blue) regions correspond to overdensities (underdensities). Dark matter halos form, grow, and merge within the overdense regions, resulting in the cosmic web of structure cosmologists observe today.

HOMEWORK Problems

Problem 1.30.1: The growth of linear dark matter overdensities.

This problem walks through a more rigorous derivation of the result $\delta \propto a$. The dynamics of the dark matter density field can be described using the following equations:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \vec{v}) = 0 \tag{1.30.11}$$

$$\frac{\partial \vec{v}}{\partial t} + (\vec{v} \cdot \nabla)\vec{v} = -\frac{\nabla p}{\rho} - \nabla \Phi \tag{1.30.12}$$

$$\nabla^2 \Phi = 4\pi G \rho, \tag{1.30.13}$$

where ρ is the dark matter density, $\vec{v}$ is its velocity, p is its pressure, and Φ is the gravitational potential. Equations 1.30.11 (the continuity equation) enforces conservation of mass, Equation 1.30.12 (the Euler equation) enforces conservation of momentum, and Equation 1.30.13 (the Poisson equation) describes how matter sources the gravitational potential. Let

$$\rho(\vec{r}, t) = \bar{\rho}(t) + \Delta\rho(\vec{r}, t) = \bar{\rho}(t)[1 + \delta(\vec{r}, t)] \tag{1.30.14}$$

$$\vec{v}(\vec{r}, t) = \vec{v}_0(\vec{r}, t) + \Delta\vec{v}(\vec{r}, t) \tag{1.30.15}$$

$$\Phi(\vec{r}, t) = \Phi_0(\vec{r}, t) + \Delta\Phi(\vec{r}, t). \tag{1.30.16}$$

By plugging these into the fluid equations and keeping terms up to first order in the fluctuations $\delta,\ \Delta\vec{v}$, and $\Delta\Phi$, combining the equations yields a differential equation for the evolution of the density contrast:

$$\ddot{\hat{\delta}} + 2H\dot{\hat{\delta}} = \left(4\pi G\bar{\rho} - \frac{c_s^2 k^2}{a^2}\right)\hat{\delta}, \tag{1.30.17}$$

where $\hat{\delta}$ is the Fourier transform of delta, k is the wavenumber associated with the Fourier transform, and $c_s^2 = \partial p / \partial \rho$ is the speed of sound. On large scales (small k), dark matter behaves as a pressureless fluid with $c_s = 0$. Thus,

$$\ddot{\hat{\delta}} + 2H\dot{\hat{\delta}} = \frac{3}{2}H^2\hat{\delta}. \tag{1.30.18}$$

a) Let $\hat{\delta} \propto t^n$ and solve for n by plugging this into Equation 1.30.18; you'll obtain two solutions for n because this is a second-order differential equation. Assume that the universe only contains matter to plug in $H(t)$.

b) Interpret both solutions obtained in part **a)** physically. Show that one of the solutions yields the expected behavior $\delta \propto a$.

Problem 1.30.2: The growth of linear baryonic overdensities.

We argued that dark matter overdensities at the time of the CMB are large enough to facilitate nonlinear structure formation by today. But why are baryonic fluctuations smaller than dark matter fluctuations at early times? The key difference is that baryons can exchange energy and momentum by interacting with each other, and are therefore *not* a pressureless fluid. Thus, for baryons, we can't drop the c_s^2 term in Equation 1.30.17.

The kinds of solutions to Equation 1.30.17 are determined by the sign of $4\pi G\bar{\rho} - \frac{c_s^2 k^2}{a^2}$. At a given time, this depends on whether k is smaller or larger than a critical wavenumber k_J; the corresponding scale $\lambda_J = 2\pi/k_J$ is referred to as the *Jeans length*.

a) Derive the Jeans length in terms of c_s, a, and $\bar{\rho}$.

b) Qualitatively describe how baryonic overdensities evolve on scales larger than and smaller than the Jeans length.

c) Before recombination, baryons feel pressure exerted by photons; the corresponding speed of sound is $c_s \approx c/\sqrt{3}$. Calculate the corresponding Jeans length assuming recombination occurs at $z = 1100$. Relate this scale to the wavenumber of the peak in the linear matter power spectrum shown in the figure below (adapted from Chabanier et al. 2019).

Problem 1.30.3: Top-hat collapse.

Consider the top-hat perturbation defined in Equation 1.30.8 A shell at radius r_i initially expands with the Hubble Flow, $v_i = H(t_i)r_i$, and thus has an initial energy per unit mass (or *specific energy*) of

$$E_i = K_i + U_i = \frac{1}{2}v_i^2 - \frac{GM(<r_i)}{r_i}, \tag{1.30.19}$$

where $M(<r_i)$ is the enclosed mass. The integrated equation of motion for this shell is

$$\frac{1}{2}\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 - \frac{GM}{r} = E. \tag{1.30.20}$$

a) Show that the system collapses (corresponding to $E_i < 0$) if $\delta_i > \Omega_m(t_i)^{-1} - 1$.

b) Show that $r(\theta) = A(1-\cos\theta)$, $t(\theta) = B(\theta - \sin\theta)$ solves the integrated equation of motion, where $\theta \in [0, 2\pi]$, $A = GM/2|E|$, and $B = GM/(2|E|)^{3/2}$.

c) Run the code snippet and qualitatively describe the plotted evolution of the system. Is this qualitatively consistent with the description near Equation 1.30.8?

```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
%config InlineBackend.figure_format='retina'

###

#parametric solution in GM = 1 units
theta_array = np.linspace(1e-1,2*np.pi-1e-1,100)

def r_tophat(theta,E):
    A = 1./(2.*np.abs(E))
    return A*(1.-np.cos(theta))

def t_tophat(theta,E):
    B = 1./(2.*np.abs(E))**(1.5)
    return B*(theta-np.sin(theta))

###

#Plot evolution of system
plt.figure(figsize=(8,6))
E_array = np.linspace(-0.5,-0.1,5)

for E in E_array:
```

run | run all | restart | restart & run all

This page titled 1.30: Structure Formation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Ethan Nadler.

1.31: Galaxy Formation

After recombination, baryons fall into the gravitational potential wells provided by dark matter halos. As hydrogen gas falls into halos, it gains kinetic energy and heats up to a temperate roughly determined by the halo mass; for a certain range of halo masses, cooling mechanisms allow the gas to cool, condense near the center of the halo, and begin the process of star and galaxy formation. The first generations of stars and galaxies flood the universe with high-energy radiation, reionizing neutral hydrogen in the intergalactic medium. A rich set of feedback processes then shape the evolution of galaxy populations over subsequent epochs.

Like nonlinear dark matter structure formation, galaxy formation and evolution are complex processes that generally require numerical simulations to model accurately. This Chapter provides a broad overview of galaxy formation theory and the connection between galaxies and dark matter halos, which are important tools for cosmological analyses based on galaxy populations.

How do Galaxies Form?

After recombination, most of the baryons in the universe are in the form of neutral hydrogen. During this epoch, dark matter overdensities are growing ($\delta \propto a$) and collapsing into dark matter halos. Consider a region in which hydrogen gas is gravitationally attracted towards the center of a halo; as it falls in, it heats up to a *virial temperature*

$$T_{\rm vir} \approx 10^5 \text{ K} \left(\frac{V_{\rm vir}}{100 \text{ km s}^{-1}}\right)^2, \tag{1.31.1}$$

where $V_{\rm vir}$ is the *virial velocity* of the halo, which is set by its mass and density profile. For example, a halo of mass $10^{12}\ M_\odot$ ($10^9\ M_\odot$) roughly has $V_{\rm vir} \approx 200 \text{ km s}^{-1}$ ($V_{\rm vir} \approx 20 \text{ km s}^{-1}$). In this way, the halo's gravitational potential sets the temperature of the infalling gas.

Gravitational acceleration isn't the only process that sets the gas temperature. In particular, hydrogen undergoes electromagnetic interactions including *collisional excitation*, *ionization*, *recombination*, and *bremsstrahlung* with photons and free electrons, which largely act to cool the gas at these early times. The net effect of these processes is captured by the *cooling function* $\Lambda(T)$, which is defined to set the characteristic cooling time via

$$t_{\rm cool} = \frac{3nk_BT}{2n_{\rm H}^2\Lambda(T)}, \tag{1.31.2}$$

where n is the total number density of gas particles and $n_{\rm H}$ is the number density of neutral hydrogen atoms in the gas. If the cooling time is shorter than the free-fall time, the gas cools more quickly than it can gravitationally equilibrate and condenses to the center of the halo.

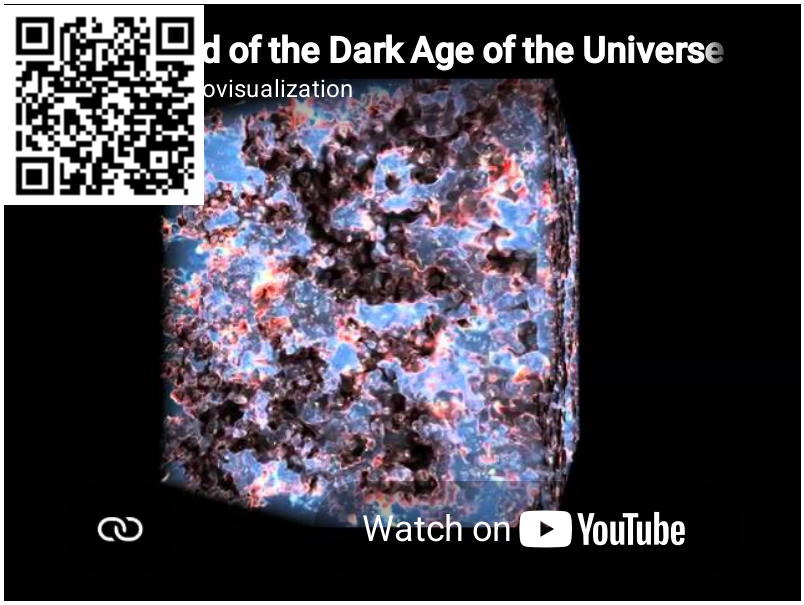
Let's estimate the minimum virial velocity necessary to efficiently cool atomic hydrogen. Other than bremsstrahlung, the cooling processes mentioned above all involve electron transitions between hydrogen energy levels, the lowest of which are separated by 13.6 eV. Thus, the gas must be warmer than roughly 10 eV to initiate these processes, corresponding to $T_{\rm vir,min} \approx 10^4$ K and $V_{\rm vir,min} \approx 10 \text{ km s}^{-1}$. This threshold is known as the *atomic cooling limit*, and it corresponds to a halo mass of roughly $10^8\ M_\odot$ today. Problem 32.1 explores how the cooling time depends on redshift, gas density, and gas composition.

Box 1.31.1

Exercise 32.1.1: Estimate the minimum halo mass necessary to efficiently cool molecular hydrogen (H_2). To convert from virial velocity to halo mass, you can combine the estimates of this relation above with $V_{\rm vir} \propto M^{1/3}$.

As a result of the cooling processes described above, sufficiently dense regions of cold gas clump into *giant molecular clouds* (GMCs) that host star formation. The dynamic range of the star formation process is remarkable: gas accretes into halos on scales of hundreds of kiloparsecs (roughly corresponding to the size of a typical halo), condenses into GMCs with characteristic sizes of tens of parsecs, and forms into stars with sizes of $\approx 10^{-7}$ parsecs. Importantly, star formation in GMCs is highly inefficient, with typical star formation timescales that are orders of magnitude larger than the corresponding free-fall times.

The first generations of stars form in dark matter halos about 100 million years after the Big Bang, transitioning the universe from post-recombination "dark ages" to the epoch of "cosmic dawn." These stars eventually die, exploding in supernovae and flooding the universe with high-energy radiation, which heats the gas in halos, inhibiting star formation. This radiation also reionizes hydrogen in the intergalactic medium, transitioning the universe from a dark, opaque neutral hydrogen gas to a transparent, ionized plasma. Several independent lines of evidence, including the optical depth of the cosmic microwave background and the absorption spectra of high-redshift quasars, indicate that the universe is fully reionized by $z \approx 6$, roughly 1 billion years after the Big Bang.



Simulation of the end of the "cosmic dark ages," the formation of the first stars and galaxies during "cosmic dawn," and the epoch of reionization. Dark regions consist of neutral hydrogen, which fills the universe after recombination. Bright regions show ionized bubbles emanating from the first generations of stars as they explode in supernovae. These ionized bubbles expand to fill the entire simulation volume by the time reionization is complete, about 1 billion years after the Big Bang.

How do Galaxies Evolve?

Early models and simulations of galaxy formation drastically overpredicted the number of galaxies as a function of galaxy luminosity, or the *galaxy luminosity function*, compared to observations. This issue, known as the "overcooling problem," is overcome by *feedback*---the effects of galaxy formation on the process itself. The supernova and photoionization heating mechanisms discussed above constitute one form of feedback. These mechanisms most strongly affect galaxies that inhabit halos with masses below $\approx 10^{12}\ M_\odot$ by heating gas to a significant fraction of these halos' virial temperatures. For halos with masses below $\approx 10^{10}\ M_\odot$, photoionization heating partially or completely shuts down subsequent star formation; we observe the relic of ancient star formation in these small halos as *dwarf galaxies* today.

The overcooling problem also applied to halos with masses above $\approx 10^{12}\ M_\odot$, which were predicted to form stars more efficiently than inferred from observations. Supernova feedback and photoionization heating do not significantly affect star formation in these halos due to their large virial temperatures. Instead, high-energy radiation from the matter orbiting supermassive black holes (SMBHs) at the centers of massive galaxies heats the surrounding gas. This mechanism, known as *active galactic nucleus (AGN) feedback*, helps bring the predicted luminosity function for the brightest galaxies into agreement with observations. Note that AGN feedback is ineffective in lower-mass systems because SMBH mass (and thus the amount of energy AGN feedback releases) decreases rapidly in smaller halos. The left panel of the figure below illustrates the resulting *galaxy--halo connection*; specifically, the relation between stellar mass and total halo mass, and the halo mass ranges in which various forms of feedback affect galaxy formation.

Galaxy properties, including (but not limited to) stellar mass, star formation rate, size, shape, color, and visual appearance (or *morphology*) evolve over cosmic time. The interplay between cooling, star formation, and feedback causes the global star formation rate in the universe to peak roughly 10 billion years after the Big Bang at $z \approx 2$, or "cosmic noon," as shown in the right panel below. Note that the rise and fall of AGN activity in the universe roughly coincides with the rise and fall of the global star formation rate.

Wechsler & Tinker 2018. *Right panel*: The average cosmic star formation rate per unit volume as a function of redshift (bottom axis) and lookback time (top axis). Points with errorbars show observational measurements, and the line shows a best-fitting star formation rate density model. Adapted from Madau & Dickinson 2014.

Galaxy morphology is historically classified according to the *Hubble sequence*, which delineates four major types of galaxies based on their visual appearance: elliptical, lenticular, spiral, and irregular. These classes of galaxies are arranged in a "tuning fork" sequence, which does **not** correspond to how galaxies typically evolve. In particular, although ellipticals and spirals are respectively referred to as "early-type" and "late-type" galaxies, modern observations indicate that the first generations of active, star-forming galaxies are predominantly spirals and irregulars. Mergers between these classes of galaxies yield ellipticals; this process often coincides with a shutdown in star formation, such that ellipticals are generally redder and less actively star-forming compared to spirals. Problem 32.2 explores how different galaxy morphologies arise and the relation between galaxy and halo evolution.

An updated version of the original Hubble sequence, which delineates four major types of galaxies: elliptical (E), lenticular (S0), spiral (S), and irregular (I). Note that galaxies do not generally evolve from left to right along the Hubble sequence. Instead, the universe is dominated by spiral and irregular galaxies at early times, which merge to form elliptical galaxies. Adapted from Kormendy & Bender 1996.

HOMEWORK Problems

Problem 1.31.1: Behavior of the cooling time.

Consider the behavior of the cooling time as a function of density, redshift, and gas composition:

a) Starting from Equation 1.31.2, argue that $t_{\text{cool}} \propto \rho_g^{-1}$, where ρ_g is the gas density. Qualitatively explain why an inverse dependence on gas density is reasonable. What does this dependence imply about the temperature profile of the gas as a function of radius from the center of a halo?

b) Use the result from part a) to show that $t_{\text{cool}} \propto a^3$. Qualitatively describe what consequences more efficient cooling at early times might have for galaxy formation.

c) Qualitatively describe how the presence of heavier elements affects the cooling time and minimum virial temperature necessary for efficient cooling. This is an important consideration because stars deposit heavier elements into the interstellar medium when they explode in supernovae.

Problem 1.31.2: Galaxy morphology and evolution.

This problem walks through a few additional aspects of galaxy morphology and evolution.

a) Using conservation of angular momentum, qualitatively describe why thin galactic disks form out of gas that falls into halos from the intergalactic medium.

b) When two spiral galaxies merge, qualitatively describe why the resulting stellar distribution is extended and elliptical.

c) Considering that small halos form first and merge together to build larger halos and how galaxy properties evolve as a result of mergers, qualitatively describe how you expect the stellar mass--halo mass relation to evolve over cosmic time.

This page titled 1.31: Galaxy Formation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Ethan Nadler.

1.32: Euclidean Geometry

One of the great privileges of teaching this class is the opportunity I have to blow your minds with a radically different understanding of the nature of space and time. The shift from a Euclidean/Newtonian understanding of space and time, to a Riemannian/Einsteinian one is centrally important to our understanding of cosmology. This chapter is entirely focused on the Euclidean geometry that is familiar to you, but reviewed in a language that may be unfamiliar. The new language will help us journey into the foreign territory of Riemannian geometry where space is curved. Our exploration of that territory will then help you to drop your pre-conceived notions about space and to begin to understand the broader possibilities -- possibilities that are not only mathematically beautiful, but that appear to be realized in the natural world.

According to Euclidean geometry, it is possible to label all space with coordinates *x*, *y*, and *z* such that the square of the distance between a point labeled by x_1, y_1, z_1 and a point labeled by x_2, y_2, z_2 is given by $(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2$. If points 1 and 2 are only infinitesimally separated, and we call the square of the distance between them $d\ell^2$, then we could write this rule, that gives the square of the distance as

$$d\ell^2 = dx^2 + dy^2 + dz^2 \tag{1.32.1}$$

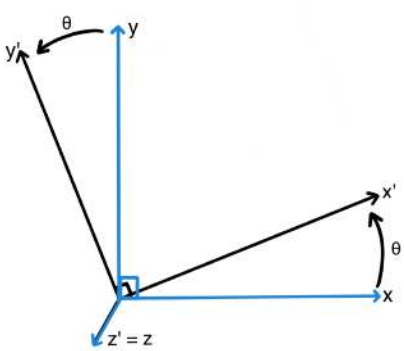
This rule has physical significance. The physical content is that if you place a ruler between these two points, and it is a good ruler, it will show a length of $d\ell = +\sqrt{d\ell^2}$. Since it is difficult to find rulers good at measuring infinitesimal lengths, we can turn this into a macroscopic rule. Imagine a string following a path parameterized by λ, from $\lambda = 0$ to $\lambda = 1$, then the length of the string is $\int_0^1 d\lambda(d\ell/d\lambda)$. That is, every infinitesimal increment $d\lambda$ corresponds to some length $d\ell$. If we add them all up, that's the length of the string.

Box 1.32.1

Exercise 1.1.1: Find the distance along a path from the origin to (x,y,z) = (1,1,1) where the path is given by

$$x(\lambda) = \lambda, y(\lambda) = \lambda, z(\lambda) = \lambda \tag{1.32.2}$$

There are many ways to label the same set of points in space. For example, we could rotate our coordinate system about the *z* axis by angle θ (with positive θ taken to be in the counterclockwise direction as viewed looking down toward the origin from positive z) to form a primed coordinate system with this transformation rule:



$$z' = z \tag{1.32.3}$$

$$y' = -x\sin\theta + y\cos\theta \tag{1.32.4}$$

$$x' = x\cos\theta + y\sin\theta \tag{1.32.5}$$

Under such a re-labeling, the distance between points 1 and 2 is unchanged. Physically, this has to be the case. All we've done is used a different labeling system. That can't affect what a ruler would tell us about the distance between any pair of points. Further, for this particular transformation, the equation that gives us the distance between infinitesimally separated points has the same form.

Figure 1: A counterclockwise rotation of the coordinate system about the z axis by θ creates a new coordinate system which we've labeled with primes. The z axis comes out of the screen and is identical to the z' axis. As is true for any point in space, point 1 can be described in either coordinate system, by specifying (x_1, y_1, z_1) or (x_1', y_1', z_1') with the relationship between the two given by the equations to the right.

Box 1.32.2

Exercise 1.2.1: Show that the distance rule of Equation 1.32.1 applied to the prime coordinates,

$$(d\ell')^2 = (dx')^2 + (dy')^2 + (dz')^2$$

gives the same distance; i.e, show that $d\ell' = d\ell$. [Warning: θ is not a coordinate here. It specifies the relationship between the coordinate systems. So, e.g., $dx' = dx\cos\theta + dy\sin\theta$.] Because this distance is invariant under rotations of the coordinate system, we call it the invariant distance.

We want to emphasize that the labels themselves, *x*, *y*, *z* or *x'*, *y'*, *z'* have no physical meaning. All physical meaning associated with the coordinates comes from an equation that tells us how to calculate distances along paths. To drive this point home, note that we could also label space with a value of *x*, *y*, *z* at every point, but do it in such a way that we would have the distance between *x*, *y*, *z* and $x+dx, y+dy, z+dz$ have a square given by

$$d\ell^2 = dx^2 + x^2\left(dy^2 + \sin^2 y dz^2\right) \tag{1.32.6}$$

For many readers, this result would look more familiar if we renamed the coordinates $r = x$, $\theta = y$, and $\phi = z$ so that we get another expression for the invariant distance,

$$d\ell^2 = dr^2 + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right) \tag{1.32.7}$$

This is the usual spherical coordinate labeling of a 3-dimensional Euclidean space by distance from origin, *r*, a latitude-like angle, θ, and a longitudinal angle, ϕ. The transformation between the two coordinate systems is given by

$$z = r\cos\theta \tag{1.32.8}$$

$$y = r\sin\theta\sin\phi \tag{1.32.9}$$

$$x = r\sin\theta\cos\phi \tag{1.32.10}$$

Box 1.32.3

Exercise 1.3.1: Show that the invariant distance given by the equation $d\ell^2 = dx^2 + dy^2$, the 2-D version of 1.32.1, and the invariant distance given by the equation $d\ell^2 = dr^2 + r^2 d\phi^2$, the 2-D version of 1.32.7, are consistent if the coordinates are related via:

$$x = r\cos\phi$$

$$y = r\sin\phi$$

Hint: use the chain rule, so that, e.g., $dx = dr\cos\phi - r\sin\phi d\phi$. (Note that the coordinate transformation equations here are obtained from the 3-dimensional case by setting $\theta = \pi/2$.)

In preparation for thinking about non-Euclidean spaces, we are going to go through how one could construct a labeling of a two-dimensional Euclidean space in polar coordinates, r, ϕ. Our construction starts with what will look like an unusual way of defining r. We define r based on the circumference of the circle rather than the distance from the origin, for reasons that will become clear later.

First we choose a center to our coordinate system. Then we label all points with r that are equidistant from that center and form a circle with circumference $C = 2\pi r$. Thus to label space with the appropriate value of r, one takes a string, ties one end down at the center, and marks out all the points that can be just reached by the other end of the string, when it is pulled straight. Then one measures the circumference of the resulting circle and labels the points on this circle with a value of r given by $r = C/(2\pi)$. We take strings of varying lengths and repeat again and again to figure out the value of r for every point in the plane.

Next, to label space with ϕ, we take one point on one of the circles and arbitrarily label that one as $\phi = 0$. We pull a string tight from the origin out to this point and beyond, and label all the points along the string with $\phi = 0$. We then march outward from the origin and when we get to a point labeled with radial value r, we make a 90° turn to the left and advance some small distance Δ. We then label this point with $\phi = \Delta/r$. Again we pull a string tight from the origin out to this point and beyond, and label all these points along the string with the same value of ϕ. We then advance another Δ around the circle and repeat, now labeling the nth iteration with $\phi = n\Delta/r$. In this manner we label all points in the space with values of ϕ. Note that when we have done this $2\pi r/\Delta$ times we will have advanced all the way around the circle (because we will have covered a distance of $2\pi r$) and the change in ϕ will be $(2\pi r/\Delta)\times\Delta/r = 2\pi$.

In a Euclidean space, such a construction leads us to the result (unproven here) that the distance $d\ell$ between two infinitesimally separated points labeled by r, ϕ and $r+dr, \phi+d\phi$ has a square given by

$$d\ell^2 = dr^2 + r^2 d\phi^2. \tag{1.32.11}$$

Note that in our construction we never made any measurement of the distance from the origin to a circle with origin as center and with circumference $C = 2\pi r$. All we know so far is the circumference of the circle. To calculate the distance from the origin to this

circle we can apply the above rule for a path that extends from the origin to the circle. Let's say a circle with circumference $C_1 = 2\pi r_1$.

Box 1.32.4

Exercise 1.4.1: Calculate the distance from the origin to the circle with circumference $C_1 = 2\pi r_1$. Do so along a path of constant ϕ using Eq. 1.32.11.

You should have got the unsurprising result that the distance from the origin to the circle with circumference $C_1 = 2\pi r_1$ is r_1. In the next chapter this will get more interesting as we examine a space for which this is *not* the case. We'll see that the distance to a circle with this circumference could be more than r_1 or less than r_1.

We constructed our coordinate system so that as θ goes from 0 to 2π at constant $r = r_1$ a distance is traversed of $2\pi r_1$. Let's now check that our rule for $d\ell$ above, Eq. 1.32.11 is consistent with this construction.

Exercise 1.4.2: Show that the parameterized path $r = r_1, \theta = \lambda$ as λ goes from 0 to 2π covers a distance of $2\pi r_1$ by integrating $d\ell$, as given by Eq. 1.32.11, along this path.

Before going on, we could take a little more care. We have shown that a particular path that takes us from the origin out to $r = r_1$ at constant ϕ has distance r_1. But how do we know this is the shortest path? Here we will demonstrate that there is not a shorter path; the one prescribed is the shortest path possible. To do so, we use a result from the calculus of variations. That result is as follows:

For $J = \int_1^2 d\mu f(q_i, \dot{q}_i, \mu)$ where $\dot{q}_i \equiv dq_i/d\mu$, the path from point 1 to 2 that extremizes J satisfies these equations

$$\frac{d}{d\mu}\left(\frac{\partial f}{\partial \dot{q}_i}\right) = \frac{\partial f}{\partial q_i} \tag{1.32.12}$$

This is a mathematical result with more than one application. In mechanics, the action is given as an integral over the Lagrangian so that

$$S = \int dt L(q_i, \dot{q}_i, t) \tag{1.32.13}$$

with $\dot{q}_i \equiv dq_i/dt$, and because a system passes from point 1 to point 2 along the path that minimizes the action, the path taken will satisfy

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) = \frac{\partial L}{\partial q_i} \tag{1.32.14}$$

which you know as the *Euler-Lagrange equations*.

In the case at hand we have length = $\int d\mu \frac{d\ell}{d\mu}$ where

$$f = \frac{d\ell}{d\mu} = \sqrt{\dot{r}^2 + r^2\dot{\phi}^2} \tag{1.32.15}$$

(note the overdot is differentiation with respect to the independent variable which here is μ again) so the shortest-length path between any two points should satisfy

$$\frac{d}{d\mu}\left(\frac{\partial f}{\partial \dot{r}}\right) = \frac{\partial f}{\partial r} \tag{1.32.16}$$

$$\frac{d}{d\mu}\left(\frac{\partial f}{\partial \dot{\phi}}\right) = \frac{\partial f}{\partial \phi} \tag{1.32.17}$$

These equations are kind of hairy, if you work them out in generality. However, we are testing to see if a particular path satisfies them, the path from the origin to $r = r_1$, and $\phi = \phi_1$ that proceeds at fixed ϕ. We could parameterize this path with $\phi = \phi_1, r(\mu)$

$= \mu r_1$ with μ running from 0 to 1. Note that $\dot{\phi} = 0$ which really simplifies the evaluation of the above equations. We will just do one term out of the first equation as an example, and leave evaluation of the rest of the terms as an exercise. In particular, we evaluate $\partial f / \partial r = (r/f)\dot{\phi}^2 = 0$.

Box 1.32.5

Exercise 1.5.1: Evaluate the three other terms ($\frac{d}{d\mu}\left(\frac{\partial f}{\partial \dot{r}}\right)$, $\frac{d}{d\mu}\left(\frac{\partial f}{\partial \dot{\phi}}\right)$ and $\frac{\partial f}{\partial \phi}$) in the two equations above, and verify that the given path does indeed satisfy these equations, thereby demonstrating that it is the shortest possible path.

Summary

1. Space can be labeled with coordinates. The same space can be labeled with a variety of coordinate systems; e.g., Cartesian or Spherical.
2. The coordinate labelings themselves have no physical meaning. Physical meaning resides in the distances between points, which one can calculate from a rule that relates infinitesimal changes in coordinates to infinitesimal distances.
3. Paths through a space can be parameterized by a single variable; we saw several examples of this.
4. The Euler-Lagrange equations can be used to prove that a particular path is (or is not) one with an extreme value of distance between a pair of points on the path. Usually the extreme is a minimum rather than a maximum.

Homework Problems

Problem 1.32.1

Starting from $d\ell^2 = dx^2 + dy^2$ prove the Pythagorean theorem that the squares of the lengths of two sides of a right triangle are equal to the square of the hypotenuse. Start off by proving it for a triangle with the right-angle vertex located at the origin, so all three vertices are at $(x, y) = (0, 0), (x_1, 0)$, and $(0, y_1)$. Be careful to use the distance rule to determine the length of each leg of the triangle, rather than your Euclidean intuition. Let's call the length of the side along the x-axis ℓ_x and similarly the other lengths ℓ_y and ℓ_h. Parameterize each path and perform the appropriate integral over the independent variable you used for the parametrization (like we did with λ in this chapter). Doing so, you should find that $\ell_h^2 = \ell_x^2 + \ell_y^2$. Having proved the Pythagorean theorem for this specially located and oriented triangle, note that since translations and rotations of the coordinate system leave our invariant distance rule unchanged, you have effectively proved it for all right triangles.

Problem 1.32.2

Prove that the hypotenuse, the straight line from $(x_1, 0)$ to $(0, y_1)$ you described in 1.1, is the shortest path between those two points.

Problem 1.32.3

Show that for a primed system that is rotated relative to the unprimed system so that

$$x' = x\cos\theta + y\sin\theta$$

$$y' = -x\sin\theta + y\cos\theta$$

the square of the invariant distance is unchanged; i.e., $dx^2 + dy^2 = (dx')^2 + (dy')^2$.

This page titled 1.32: Euclidean Geometry is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.33: Curvature

We introduce the notion of "curvature" in an attempt to loosen up your understanding of the nature of space, to have you better prepared to think about the expansion of space.

About 2300 years ago Euclid laid down the foundations of geometry with 23 definitions and five postulates. This was all done in an attempt to capture what were obvious fundamental properties of space, so that they could be used as starting points to prove other things about space -- such as that the sum of the angles of a triangle is 180 degrees, and that the ratio of a circle's circumference to diameter is the same for all circles. The first four postulates could be put rather succinctly. The fifth one though is a bit unwieldy. In translation from the original Greek, taken from Wikipedia, we have [in two-dimensional geometry]:

> *If a line segment intersects two straight lines forming two interior angles on the same side that sum to less than two right angles, then the two lines, if extended indefinitely, meet on that side on which the angles sum to less than two right angles.*

There were many attempts to prove this fifth postulate, also known as the "parallel postulate" from the other four. Eventually it was realized that it can't be done.

We have also come to *experimentally* determine that space is different from what one gets with all five of Euclid's axioms together. Our theoretical understanding of the nature of space (consistent with all measurements) allows for the possibility that two lines as described in the fifth postulate might never meet, on either side of the line they both intersect. Throwing out this fifth postulate leads to the possibility that the angles of a triangle sum up to something different from 180 degrees, and that the ratio of circumference to diameter of a circle can vary depending on location of the circle center and the size of the circle.

The ratio of circumference to diameter that is different from π is a signature of a property called "curvature." Euclidean geometry is a geometry with zero curvature. In this section we will study both two and three-dimensional curved (and therefore "non-Euclidean") spaces. One way we will gain some intuition about these spaces is by *embedding* them in a Euclidean space of one extra dimension. We will emphasize here that there need not be any physical reality to the extra-dimensional space. We can describe curved spaces mathematically without any reference to extra dimensions. Space can "curve" without having to "curve into" any external space. Coming to terms with this idea will, perhaps, make you more comfortable with the idea that space can expand without expanding into anything else.

Let's get started.

We can label space with coordinates, for example, we could label every point in a 2-dimensional space with an x value and a y value. These coordinates are just labels, with no physical meaning, until we also say something about the distance between infinitesimally separated pairs of points. For example, in a two-dimensional Euclidean space with which you are familiar, the square of the distance between x, y and $x + dx, y + dy$ is given by:

$$d\ell^2 = dx^2 + dy^2 \tag{1.33.1}$$

The physical interpretation of $d\ell^2$ is as follows:
The length of a ruler with an end on each of the two points is $\sqrt{d\ell^2}$.

Expressions for the distance are not always as simple as the one above. Even for the same physical space, with different coordinate schemes the equation for $d\ell^2$ will look different. For example, we could choose polar coordinates instead of Cartesian ones and then distances would be given by:

$$d\ell^2 = dr^2 + r^2 d\phi^2 \tag{1.33.2}$$

This is the same space, just described with different coordinates.

Note that one can find the distance along any path through the space by calculating $\int d\ell$ along the path.

The space we are describing in this chapter so far (and its higher- and lower-dimensional versions) we call "Euclidean" because these spaces are consistent with geometry as described by Euclid. It turns out though that space is ***not*** Euclidean (although a Euclidean description is often a very good approximation). This is a bit startling. If you are not startled by it, you don't understand it yet. But don't worry; you will. And hopefully your mind will be blown.

From Einstein we learned that space can be quite different from Euclidean. To begin to free your mind from its Euclidean constraints, let's consider a non-Euclidean space that we label with r and ϕ using the same construction as we outlined in the previous chapter. The construction is exactly the same, but the strange nature of the space is revealed by this different rule for the distance, $d\ell$, between r, ϕ and $r + dr, \phi + d\phi$:

$$d\ell^2 = \frac{dr^2}{1 - kr^2} + r^2 d\phi^2 \tag{1.33.3}$$

At the moment, the introduction of the $1 - kr^2$ factor should just look arbitrary. We will eventually derive this distance rule from an assumptions of homogeneity and isotropy of the space. For now, let's explore its geometrical implications.

For the three following boxes (four exercises) assume the space is one governed by the distance rule of Eq. 1.33.3.

Box 1.33.1

Exercise 2.1.1: How long would a path be that stretches from $r = 0$ to $r = r_1$ at constant ϕ? Call the length ℓ and express it as an integral that depends on r_1 and k. Assume that r_1 is much less than $\sqrt{1/k}$. Make a Taylor expansion that approximates the integrand so that it contains the first order corrections due to $k \neq 0$. After this approximation, do the integral. [Hint: many students in my experience have trouble with this Taylor expansion, hence this hint. Use this first order Taylor expansion result: $(1 + \epsilon)^n = 1 + n\epsilon$ where ϵ is some small number and apply it to the term $\sqrt{1/(1 - kr^2)}$.]

Box 1.33.2

Exercise 2.2.1: Consider the set of points all at $r = r_1$ with all values of ϕ. Is this a circle? What is the circumference of this object as a function of its radius, ℓ? Recall that because of how the coordinate system was constructed, you can assume that r, ϕ and $r, \phi + 2\pi$ are the same point. First find the circumference as a function of r_1 and then use your result from the previous problem to express it as a function of k and ℓ. [Don't get too hung up on solving for ℓ as a function of r_1. If you take advantage of some additional appropriate approximations the algebra is not too bad, but don't spend too much time trying to figure it out.]

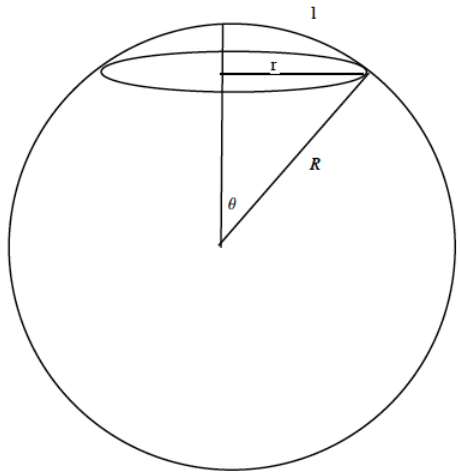
Exercise 2.2.2: Discuss the result from Exercises 2.1.1 and 2.2.1 and what it means qualitatively for the circumference-radius relationship for circles in spaces with $k < 0$, $k > 0$ and $k = 0$. [Note that you can do this even if you did not manage to get ℓ as a function of r in the above exercise.]

Box 1.33.3

Exercise 2.3.1: If you were a two-dimensional creature, and could travel around this space with a measuring tape, describe in a few sentences at least one way for measuring the value of k.

Embedding

Sometimes it is possible to visualize a non-Euclidean space, such as the one with invariant distance rule given by Equation 1.33.3, by embedding it in a higher-dimensional Euclidean space. Such an embedding, into a space with one extra dimension, is shown in the figure for $k > 0$. When using such an embedding diagram it is important to keep in mind that there are extra dimensions in the diagram whose sole purpose is visualization -- they have no physical significance. In the figure here, the space we are describing is the two-dimensional sphere; the radial dimension is fictional, included here only for purposes of visualization.



Note that every point on the sphere can be labeled by the latitude-like coordinate θ and the longitudinal coordinate (not shown) ϕ. Also, at every point one can convert θ and ϕ to r and ϕ. In a homework problem you will derive Equation 1.33.3 starting from the assumption that the three-dimensional space used for the embedding is Euclidean.

Box 1.33.4

Exercise 2.4.1: Observe the circle at constant coordinate value r in the embedding diagram. The distance (traveled on a path restricted to the 2-dimensional space of the sphere), from the origin (top of the sphere) to any part of the circle, in the 2-dimensional space of the sphere, is ℓ. Is the circumference of the circle greater than, equal to, or less than $2\pi\ell$? Compare to the relevant result in the boxes above.

Not all spaces can be embedded by placing them in just one extra dimension. For example, the $k<0$ space requires two extra dimensions for embedding. *Part* of the $k<0$ space is sometimes shown embedded in just one extra spatial dimension, with the two-dimensional surface having a shape similar to a saddle. One can start to see the problem here because if the diagram were extended, the saddle would curve into itself. Such self intersection can only be avoided by introduction of yet another fictional extra dimension.

Three-dimensional Homogeneous and Isotropic Spaces

Let us now take things up one dimension into 3-D.

Previously we asserted that one could label a 3-dimensional Euclidean space with coordinates r, θ, and ϕ such that points separated by dr, $d\theta$, and $d\phi$ would be separated by a distance (as one would measure with a ruler) with square given by

$$d\ell^2 = dr^2 + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right) \tag{1.33.4}$$

A space that can be labeled in this way is homogeneous (invariant under translations) and isotropic (invariant under rotations). The easiest way to see this is to remember that there's a coordinate transformation to Cartesian coordinates for which

$$d\ell^2 = dx^2 + dy^2 + dz^2 \tag{1.33.5}$$

Now the homogeneity is more evident, since transforming x to $x' = x + L$ would clearly leave the distance rule unchanged. We've also already seen that rotations leave the distance rule unchanged. So, the space is homogeneous and isotropic. If we choose to label it with spherical coordinates about a particular origin, our labeling obscures the homogeneity and isotropy, but the space itself is still homogeneous and isotropic.

It turns out that whether one can label space in this way or not is a matter to be settled by experiment. It's not necessarily true. Even if we restrict ourselves to completely homogeneous and isotropic geometries, we can mathematically describe spaces that cannot be labeled in this way.

What *is* generally true is that all three-dimensional homogeneous and isotropic spaces can be labeled with coordinates r, θ, and ϕ such that

$$d\ell^2 = \frac{dr^2}{1-kr^2} + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right) \tag{1.33.6}$$

for k a constant that can be positive, negative or zero. Euclidean space is a special case with $k = 0$.

BOX 1.33.5

Exercise 2.5.1: You know that in a Euclidean space the relationship between radius and area of a sphere is $A = 4\pi r^2$ with r specifying the radius. Note that the angular ($d\phi$ and $d\theta$) parts of the invariant distance equation are unchanged by having $k \neq 0$. Therefore we still have $A = 4\pi r^2$ even if $k \neq 0$. I also claim that the relationship between sphere area and radius does depend on k. How can these statements both be true?

We can construct the homogeneous and isotropic three-dimensional space and derive its invariant distance rule, at least for the case of $k > 0$, by embedding it in a 4-dimensional Euclidean space. In a 4-dimensional Euclidean space we can have a coordinate system consisting of three dimensions x, y, z, that are all orthogonal to each other, and a fourth we will call w that is orthogonal to each of the $x, y,$ and z directions. Impossible as this is to visualize, we can describe it mathematically. The distance between w, x, y, z and $w + dw, x + dx, y + dy, z + dz$ is given by

$$d\ell^2 = dw^2 + dx^2 + dy^2 + dz^2. \tag{1.33.7}$$

In this 4-dimensional space, we construct a three-dimensional subspace that is the set of points all the same distance, R, from a common center. Let's center it on the origin so our subspace satisfies this constraint:

$$w^2 + x^2 + y^2 + z^2 = R^2. \tag{1.33.8}$$

This subspace is homogeneous (all points are the same) and isotropic (all directions are the same). You can see that this is true by imagining it's two-dimensional analog, a sphere, which is the set of all points satisfying $x^2 + y^2 + z^2 = R^2$.

It will be helpful at this point to swap out the Cartesian x, y, z for the spherical coordinate system r, θ, ϕ so we have

$$d\ell^2 = dw^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2) \tag{1.33.9}$$

and our constraint equation can be written as

$$w^2 + r^2 = R^2. \tag{1.33.10}$$

From this new version of the constraint equation, we can see that if r changes by some amount then we will necessarily have to have a change in w in order to continue to satisfy the constraint. The exact relationship between differential changes you can easily work out to be $2wdw + 2rdr = 0$ (because changing r by dr ends up changing r^2 by $2rdr$ and likewise for w and dw and since R is fixed $dR^2 = 0$). Using this relationship to eliminate dw^2 from our invariant distance expression, and using the constraint equation to eliminate w^2 in favor of r^2 and R^2 we get

$$d\ell^2 = \frac{dr^2}{1 - r^2/R^2} + r^2(d\theta^2 + \sin^2\theta d\phi^2). \tag{1.33.11}$$

We see that our subspace has an invariant distance expression of the form we were intending to derive, and it is exactly the one introduced above if we make the identification $R^2 = 1/k$.

Summary

Let us now summarize the key points in our study of spatial geometry. First, the most general, high-level points:

1. We can label the continuum of points in a space with coordinates.
2. These coordinates have no physical meaning on their own.
3. Physical meaning comes from the combination of the coordinates and a rule for converting infinitesimal coordinate differences into an infinitesimal distance.

Now let's summarize at a greater level of detail:

1. We studied two-dimensional and three-dimensional homogeneous and isotropic spaces. Homogeneity means the spaces are the same everywhere; e.g., no matter where one is in the space, one would find the same relationship between circumference and radius of circles. Isotropic means that there are no special directions in the space.

2. We labeled such spaces with polar (for 2D) and spherical (for 3D) coordinates. We used an explicit construction for the 2D case. We defined the radial coordinate as a ``circumferential coordinate'' that gives a circumference for a circle as equal to $2\pi r$ by definition. We defined the angular coordinate ϕ, in the 2D case, so that the distance from r, ϕ to $r, \phi + d\phi$ is equal to $rd\phi$. Since the circumference is $2\pi r$ this means r, ϕ and $r, \phi + 2\pi$ are the same point.
3. We first asserted that a 2D homogeneous and isotropic space with coordinates constructed as above has the following distance rule: $d\ell^2 = dr^2/(1-kr^2) + r^2 d\phi^2$, where k is some constant with units of inverse area. We also asserted that in the 3D case the distance rule in spherical coordinates (with an analogous construction procedure that we did not explicitly describe) is $d\ell^2 = dr^2/(1-kr^2) + r^2(d\theta^2 + \sin^2\theta d\phi^2)$.
4. Partly to demonstrate that these are indeed homogeneous and isotropic spaces, we then demonstrated how such a space could be constructed, for the case of $k > 0$, by *embedding* it in a Euclidean space with one additional dimension. We were able to visualize such an embedding for the 2D case. Although we could not visualize the embedding in the 3D case, we at least mathematically treated its embedding in a 4D Euclidean space to derive the distance rule.

We also introduced techniques for doing geometrical calculations. We introduced paths through a space described by use of an independent parameter; e.g., $x(\lambda) = \lambda, y(\lambda) = \lambda$ with λ ranging from 0 to 1. If these are Cartesian coordinates in a 2-dimensional Euclidean space then this describes a straight line from $(0,0)$ to $(1,1)$. We calculated distances along such parameterized paths by calculating $\int ds = \int d\lambda (ds/d\lambda)$ from one endpoint of the path to another. Continuing with our example, that's length = $\int_0^1 d\lambda \sqrt{(dx/d\lambda)^2 + (dy/d\lambda)^2} = \int_0^1 d\lambda \sqrt{1+1} = \sqrt{2}$. We also reminded you of a result from the calculus of variations, that you have seen in your study of classical mechanics, and applied it to such length integrals, in order to derive the differential equations obeyed by *extreme* paths, which are usually (but not always) the shortest paths.

We used these calculational techniques to calculate circumferences and radiuses of circles in the 2D case. With our exploration, in this way, of homogeneous and isotropic spaces, our fervent desire is that your mind is now freed somewhat from its Euclidean intuitions about the nature of space. If you can imagine that the area interior to a circle can be larger than πr^2, where $r \equiv C/(2\pi)$ and C is the circumference, or the volume interior to a sphere can be larger than $4/3\pi r^3$ where $r^2 = A/(4\pi)$ and A is the surface area, then you are now better prepared to contemplate the expansion of space.

The expansion of space happens over time. In the next two chapters we thus turn our attention to the geometry of spacetime. We'll find in Chapter 4 that spatial distances are no longer invariant under changes of coordinate systems; you may have been exposed to this before via the phenomenon of Lorentz contraction. We'll therefore be forced to alter how we relate coordinate differences between infinitesimally separated pairs of points to things we can actually measure. In particular, for this purpose, we will introduce the ``invariant distance'' that presumably you have studied in special relativity.

HOMEWORK Problems

For the homework problems we will be considering a two-dimensional space labeled with coordinates r and ϕ with invariant distance given by

$$d\ell^2 = \frac{dr^2}{1-kr^2} + r^2 d\phi^2$$

In the following, assume $k > 0$ unless otherwise specified.

Problem 1.33.1

How long is the path that runs from $r = 0$ to $r = r_1$ at constant ϕ? Call the length ℓ and express it as a function of r_1 and k. Assume that $r_1 < \sqrt{1/k}$. Unlike in the exercises, do not use a Taylor series approximation to $1/(1-kr^2)$.

Problem 1.33.2

Consider the set of points all at $r = r_1$ with all values of ϕ. This set is a circle since it is the set of all points located a particular distance away from another point ($r = 0$). What is the circumference of the circle? First find the circumference as a function of r_1 and then use your result from the previous problem to express it as a function of k and ℓ.

Problem 1.33.3

If you were a two-dimensional creature, and could travel around this space with measuring tape, describe in a few sentences at least
one way for measuring the value of k.

Problem 1.33.4

Consider the embedding diagram in the chapter. Keep in mind that the 3-dimensional space, in which the 2-dimensional sphere is embedded, is Euclidean. Use what you know about Euclidean geometry to show that the square of the length of the path between r, ϕ and $r+dr, \phi+d\phi$, *constrained to lie in the sphere*, is indeed given by $\frac{dr^2}{1-kr^2}+r^2d\phi^2$ for the appropriate choice of k as a function of the radius of the sphere R. Also specify that function.

Don't let this calculation lead you astray conceptually. Although it's perfectly fine, and a useful tool for visualization, to consider a two-dimensional space embedded in a three-dimensional Euclidean space, one can have curved two-dimensional spaces without there needing to be a third dimension. We see this mathematically in this chapter as we can do things like calculate observables (lengths) without any reference to an additional (third) dimension.

Problem 1.33.5

Fun with the Schwarzschild solution. The space outside of a central spherically symmetric mass distribution can be labeled with coordinates r, θ, ϕ such that

$$d\ell^2=(1-r_s/r)^{-1}dr^2+r^2\left(d\theta^2+\sin^2\theta d\phi^2\right)$$

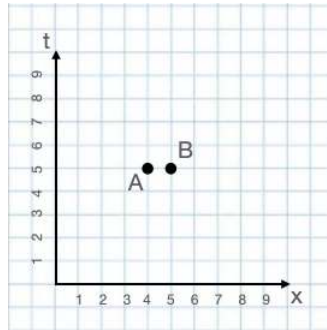
where $r_s=2GM/c^2$ is the "Schwarzschild radius" and M is the total mass of the mass distribution. The azimuthal angle ϕ runs from 0 to 2π. For the Earth, r_s is about 9 mm.

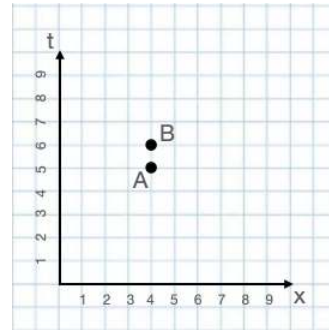
Consider two concentric circles in the same plane with centers at the center of the mass distribution ($r=0$). If the spatial geometry were Euclidean, the differences in their circumferences would be $\Delta C=2\pi\Delta\ell$ where $\Delta\ell$ is the radial distance between the two circles. But the presence of the mass means the spatial geometry is not Euclidean, and instead given by the Schwarzschild solution. If the two circles are at coordinate values r_1 and r_2, show that, for r_s much smaller than r_1 or r_2 one instead gets $\Delta C=2\pi\Delta\ell\ -\pi r_s\ln(r_2/r_1)$. Hints: 1) Choose the circles so that it's just ϕ that's changing, with θ fixed to $\pi/2$ and 2) Taylor expand to first order so $\sqrt{(1-r_s/r)^{-1}}=1+\frac{1}{2}\frac{r_s}{r}$.

Note that if we take r_2 to be 42,000 km (about the distance to geostationary orbit from the center of the Earth) and r_1 to be 6,000 km (the distance from center of Earth to the surface), the correction to the difference in the circumferences is $\pi r_s\ln(r_2/r_1)$= 11 cm. A very small correction! The spatial geometry around Earth is very close to Euclidean.

This page titled 1.33: Curvature is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.34: Galilean Relativity





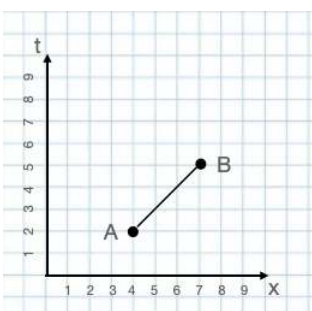
We now extend our discussion of spatial geometry to spacetime geometry. We begin with Galilean relativity, which we will then generalize in the next section to Einstein (or Lorentz) relativity.

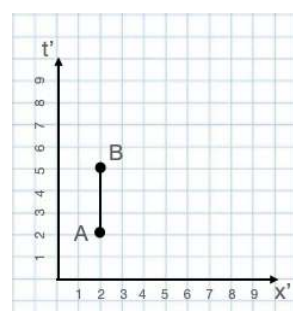
The notion of relativity of motion is not something new with Einstein. It's built into Newtonian mechanics as well. The idea is that the results of experiments done in any inertial frame will be the same; i.e., one will not be able to determine by some experimental means what frame is at "absolute rest" and which frames are moving -- all motion is relative. An inertial frame is one in which Newton's laws of motion are satisfied. These laws prescribe accelerations, not velocities, so a frame that is moving at constant relative velocity with respect to an inertial frame must itself be an inertial frame.

Let's begin by considering a labeling of all points in space, at all points in time, with t, x, y, z. Further, let's assume the spatial part of this is a Cartesian coordinate system such that a ruler (at rest in this coordinate systems) with one end at t, x, y, z and the other at $t, x + dx, y + dy, z + dz$ has length whose square is $d\ell^2 = dx^2 + dy^2 + dz^2$ and the time labeling is such that a clock that goes from t, x, y, z to $t + dt, x, y, z$ will indicate that the time elapsed is dt. Further, we will assume that this reference frame is inertial: Newton's laws of motion are satisfied in this frame.

An example of such a coordinate system, although with just one spatial dimension, is shown in the first two figures in this chapter. A ruler stretching from point A to point B in the figure with these points horizontally displaced from each other, if the units of the tick marks on the x axis are in cm, would measure 1 cm. A clock that moves from A to B in the second figure, if the units on the t axis are seconds, would show a time elapsed of one second.

Now imagine a ball as it travels at uniform velocity from a spacetime point A to a spacetime point B. Let's set this up so that point A is $t = 2$ seconds and $x = 4$ cm and point B is at $t = 5$ seconds and $x = 7$ cm as in the figure to the left. From this given information we can conclude that the ball is moving toward increasing values of x at constant speed $v = 1$ cm/s. Let's define another reference frame, called the primed frame, which is coincident with the original unprimed frame at $t = t' = 0$ but relative to the unprimed frame is moving to the right at speed $v = 1$ cm/s just like the ball. In the primed frame (see the figure to the right) the origin of the primed frame is not moving (by definition), the ball is not moving, and the origin of the unprimed frame is moving toward decreasing values of x' at speed v. The principle of relativity assures us that we can construct this primed frame so that it also has the same nice properties of the unprimed frame; i.e., the Pythagorean theorem applies for lengths, and the readings of stationary clocks make sense; i.e., a clock that goes from t', x', y', z' to $t' + dt', x', y', z'$ will indicate that the time elapsed is dt'. The principle of relativity also leads us to the conclusion that if the unprimed frame is an inertial frame, the primed frame will be as well.





So far we have not mentioned Galileo. The above statements all follow from the principle of relativity, and our assumption that space is Euclidean. We specialize to Galilean relativity when we assume that the relationship between these two coordinate systems is:

$$\begin{aligned} t' &= t, \\ x' &= x - vt, \\ y' &= y, \text{ and} \\ z' &= z \end{aligned} \tag{1.34.1}$$

Because the relationship is time dependent (and because we are anticipating the generalization to Lorentz transformations), we have explicitly included time in the transformation, even though time transforms trivially. We will call such a transformation a "Galilean Boost." In this case the above equations describe the relationship between a reference frame we denote without primes (the unprimed reference frame) and a reference frame that we denote with primes (the primed reference frame), where the primed reference frame is moving relative to the unprimed reference frame with speed v in the $+x$ direction.

One can quickly verify that, if evaluated at the same time t, the distance between two infinitesimally separated points is unchanged by this transformation. Further, the transformation is symmetric. That is, there is nothing special about the primed frame relative to the unprimed frame. The reverse transformation, found by solving the above for t, x, y, and z is

$$\begin{aligned} t &= t', \\ x &= x' + vt', \\ y &= y', \text{ and} \\ z &= z' \end{aligned} \tag{1.34.2}$$

which is the same equation as above except with the replacement of v with $-v$. That is, it is exactly the same rule, with $v \to -v$ because while the primed frame is moving relative to the unprimed frame toward higher x, the unprimed frame is moving relative to the primed frame toward lower x'.

Box 1.34.1

Exercise 3.1.1: With the Galilean boost transformation, velocities add in a simple manner. If $u' = dx'/dt'$ where $x'(t')$ is the x' location of some particle at time t', find $u = dx/dt$ as a function of u' and v.

Box 1.34.2

Exercise 3.2.1: Show that Newton's Law for a spring is invariant under a Galilean transformation. In particular, show that the equation $m\ddot{x} = -k(x - x_c)$ is invariant under a Galilean transformation assuming the location of the force-free point of the spring, x_c, is transformed as well. Here we have taken the spring equilibrium length to be zero and the spring constant to be k.

Box 1.34.3

Exercise 3.3.1: Discuss what this invariance means for applicability of Newton's laws in both the primed and unprimed frames, and the question: "is such a law consistent with the principal of Galilean relativity?"

This page titled 1.34: Galilean Relativity is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.35: Einstein Relativity

In the 19th century it was discovered that the Maxwell Equations describing electric and magnetic fields, a grand synthesis of the results of many different experiments, unlike Newton's laws of motion, are not consistent with Galilean relativity. A priori, the solution was not clear. One possible reason for this inconsistency, taken seriously at the time, was that the principle of relativity is wrong; i.e., there actually is an absolute rest frame, and our motion could be detected with respect to it with the appropriate experiment. Indeed, there was a significant experimental program to detect our motion with respect to absolute rest defined by a medium called "the ether."

Another possibility, is that while the principle of relativity holds, its specific implementation as Galilean relativity does not. As you know, because you have studied special relativity, this is indeed the correct solution to the puzzle of the Maxwell Equations lack of invariance under a Galilean transformation.

It turns out that the "Galilean Boost" can be generalized to a "Lorentz Boost" that is also consistent with the principle of relativity. The primed and unprimed coordinate systems constructed as before, under a Lorentz boost are related as:

$$\begin{aligned} t' &= \gamma(t - vx/c^2) \\ x' &= \gamma(x - vt), \\ y' &= y, \text{ and} \\ z' &= z \end{aligned} \tag{1.35.1}$$

where $\gamma \equiv 1/\sqrt{1 - v^2/c^2}$. In the limit that $c \to \infty$ this reduces to the Galilean boost. As can be easily shown (see the homework problem) the reverse transformation is the same rule with $v \to -v$. Most importantly, the Maxwell equations *are* invariant under this transformation.

One of the more spectacular consequences of the Maxwell Equations is that one of their solutions is waves traveling at the speed of light. If the Maxwell equations are correct in all inertial frames, then this implies that these waves will be moving at the speed of light in all inertial frames. To your Galilean intuition this is quite startling as it violates the simple rule for addition of velocities you derived in the previous chapter.

The result can be easily demonstrated from the Lorentz transformation above. Here we sketch out the process, and you can fill in the details by performing the exercise that follows. Imagine a particle traveling at the speed of light. Let's parameterize its path through spacetime with the independent variable λ so that $t = \lambda$ and $x(\lambda) = c\lambda$. Then we have (by direct substitution into the Lorentz transformation) that $t' = (\gamma/c)(c - v)\lambda$ and $x' = \gamma(c - v)\lambda$. The speed of this particle in the primed frame is

$$\frac{dx'}{dt'} = \frac{dx'}{d\lambda}\frac{d\lambda}{dt'} = \frac{dx'}{d\lambda}\left(\frac{dt'}{d\lambda}\right)^{-1} = c. \tag{1.35.2}$$

Thus we see the Lorentz transformation tells us that a particle traveling at speed c in one frame will be traveling at speed c in another. This result is consistent with our claim that the Maxwell equations are invariant under the Lorentz transformation, since a consequence of the Maxwell Equations is that electromagnetic waves travel at speed c.

Box 1.35.1

Exercise: 4.1.1: Fill in the steps in the above derivation.

Unlike rotational coordinate transformations that preserve spatial distances between pairs of points, a Lorentz transformation does not. The spatial separation between (x, t) and $(x + dx, t)$ is dx. The spatial separation between these points in the prime frame is γdx, as one can see from the transformation rule. How can length depend on reference frame? Key to resolving this apparent paradox is the fact that in the primed frame the two events are not simultaneous. We won't sort out these apparent paradoxes here.

We will, however, introduce a quantity that, unlike spatial length, *is* invariant under Lorentz transformations. For Cartesian spatial coordinates, the square of the invariant distance between event (t, x, y, z) and event $(t + dt, x + dx, y + dy, z + dz)$ is given by

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2. \tag{1.35.3}$$

This quantity has the following two-part physical interpretation:

1. For $ds^2 > 0$, $\sqrt{ds^2}$ is the length of a ruler that connects the two events and is at rest in the frame in which the two events are simultaneous.

2. For $ds^2 < 0$, $\sqrt{-ds^2/c^2}$ is the time elapsed on a clock that moves between the two events with no acceleration.

Why is this quantity invariant under boosts? That's a deep question, and I'm not sure we have the fullest possible answer yet sorted out. We do know that the Maxwell equations are a synthesis from experiments, their form is invariant under a Lorentz transformation, and the Lorentz transformation preserves the invariant distance.

Let's show that the invariant distance is indeed invariant under a Lorentz transformation. For specificity, take it to be the transformation appropriate for a boost in the $+x$ direction with speed v. For simplicity, we will take your two coordinate systems to be coincident at their origins (i.e. $t = x = y = z = 0$ is the same point as $t' = x' = y' = z' = 0$), use the origin as one point, let's call it point A, and $t = dt, x = dx, y = dy, z = dz$ as the other, let's call it point B. The invariant distance between these two points is

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2. \tag{1.35.4}$$

In the primed frame point A is the origin (by construction) and point B is labeled with $t' = dt', x' = dx', y' = dy', z' = dz'$.

Thus the invariant distance in the primed frame is

$$(ds')^2 = -c^2(dt')^2 + (dx')^2 + (dy')^2 + (dz')^2. \tag{1.35.5}$$

We want to show that $(ds')^2 = ds^2$.

So we need to know how dt' and dx' etc., are related to dt, dx, etc. Since the boost is in the x direction by speed v we have, from Equation ???:

$$\begin{aligned} dt' &= \gamma(dt - vdx/c^2) \\ dx' &= \gamma(dx - vdt), \\ dy' &= dy, \text{ and} \\ dz' &= dz. \end{aligned} \tag{1.35.6}$$

The easier direction to go here is to start with

$(ds')^2 = -c^2(dt')^2 + (dx')^2 + (dy')^2 + (dz')^2$ and work our way toward showing that this equals $-c^2dt^2 + dx^2 + dy^2 + dz^2$ so let's do that. By completing the exercise below you will see that these are indeed equivalent, and thereby show that the invariant distance between two infinitesimally separated points in spacetime is invariant under a Lorentz transformation.

Box 1.35.2

Exercise 4.2.1:

Show that $(ds')^2 = -c^2(dt')^2 + (dx')^2 + (dy')^2 + (dz')^2$ is the same as $-c^2dt^2 + dx^2 + dy^2 + dz^2$ if the two coordinate systems are related by a Lorentz transformation; i.e., complete the demonstration as set up in the text immediately above. To give you a little less writing to do, feel free to drop the y and z coordinates (and their primed versions) since their transformation is trivial. Keep in mind that $1/\gamma^2 = 1 - v^2/c^2$.

Rather than the Lorentz transformation itself, the key thing to take away from this chapter is the definition of the invariant distance. We will be using it for the rest of the course, generalized to spacetimes with "curvature." Before doing so, we give some exercises here in which you get to make use of the invariant distance to solve problems in the more familiar context of a flat spacetime, the so-called Minkowski space you are familiar with from special relativity. A Minkowski space is simply a spacetime that can be labeled with t, x, y, z such that the invariant distance is given by Equation 1.35.3.

In Minkowski space, as one of the homework problems asks you to show, a finite (as opposed to infinitesmial) version of the invariant distance equation is also true:

$$(\Delta s)^2 = -c^2(\Delta t)^2 + (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 \tag{1.35.7}$$

for trajectories that are straight lines, with $\Delta s \equiv \int d\lambda \frac{ds}{d\lambda}$ also invariant under Lorentz transformations.

To demonstrate some of the utility of the invariant distance equation let's use it to derive the phenomenon of time dilation. Specifically, let's calculate the time that elapses on a clock traveling in a straight line at speed v from x_1, t_1 to x_2, t_2. We will find that it is *not* $t_2 - t_1$.

First, let's draw the trajectory of the clock on two spacetime diagrams: one with a coordinate system in which the clock is moving with speed v and the other with a coordinate system that has the clock at rest.

The time that elapses on the clock as it travels between these two points will be $t_2' - t_1'$. We know this because we assume it is a good clock and the primed coordinate system has been constructed so that this is the case for clocks at rest in the primed coordinate system. (We also assume the same about the construction of the unprimed coordinate system, that a clock at rest there, will show a time elapsed of $t_2 - t_1$ as it travels, without spatial translation, from a spacetime point with time coordinate t_1 to a spacetime point with time coordinate t_2).

For the first coordinate system we have an invariant distance between points 1 and 2:

$$(\Delta s)^2 = -c^2(t_2 - t_1)^2 + (x_2 - x_1)^2 = -c^2(\Delta t)^2 + (\Delta x)^2.$$

For the prime system we have an invariant distance between points 1 and 2:

$$(\Delta s')^2 = -c^2(t_2' - t_1')^2.$$

Now we set them equal and solve for $t_2' - t_1'$:

$$-c^2(t_2' - t_1')^2 = -c^2(\Delta t)^2 + (\Delta x)^2$$

$$(t_2' - t_1')^2 = (\Delta t)^2 - \frac{(\Delta x)^2}{c^2}$$

$$\text{note that, } \frac{(\Delta x)^2}{(\Delta t)^2} = v^2$$

$$(t_2' - t_1')^2 = (\Delta t)^2\left(1 - \frac{v^2}{c^2}\right)$$

$$t_2' - t_1' = \gamma^{-1}\Delta t$$

So we see that the time elapsed on the clock is not $\Delta t = t_2 - t_1$ but instead $\Delta t/\gamma$.

Note that we could also have just calculated $(\Delta s)^2$ in the unprimed frame and used our physical interpretation of $\sqrt{-(\Delta s)^2/c^2}$ (for $(\Delta s)^2 < 0$) as the time that elapses on a clock traveling from point 1 to point 2.

Box 1.35.3

Exercise 4.3.1: In the above derivation we identified $\Delta x/\Delta t$ as the speed of the clock v. Why is this justified?

Summary and Discussion

In our study of spatial geometry we came to view coordinates as mere labelings of points in a space, with no physical significance on their own. Physical significance came through a rule that related infinitesimal differences in the coordinate values of a pair of points to an infinitesimal distance, ds. Now we have extended space to spacetime with the introduction of a temporal coordinate, that so far we have always called t. Once again we label all the points in a spacetime with coordinates that have no physical significance on their own. Physical meaning comes through a rule relating infinitesimal differences in the coordinate values of a pair of points to something observable, although in this case the rule is a bit more complicated.

You might ask why the rule is more complicated; why is it the way it is? To which one could answer, "we do not know; we just work here!" To expand on this, we do not know why the rule is as it is, but we have discovered through experiments, and abstraction from these experiments, that this rule works. Through the exercises and homework problems you can see how this rule naturally incorporates the phenomenon of time dilation. You could also work out for yourself how it incorporates the phenomenon of Lorentz contraction. The rule basically encapsulates what we have *discovered* about the local structure of spacetime.

The rule is telling us about the local structure of spacetime in the sense that it is telling us about time spans and lengths associated with points that are only separated by infinitesimal amounts. This local aspect is preserved as we move on, in subsequent chapters, to study the global structure. Thus the infinitesimal invariant distance is an important concept in our study of the expansion of the universe.

In the table below we summarize the physical meaning of ds^2 and the rules relating coordinate separations to observables for a variety of spaces/spacetimes and coordinate systems. So far the spacetimes we have introduced have been static; not evolving in time. In the next chapter we will introduce a dynamic spacetime, a 1+1-dimensional analog of the expanding spacetime we appear to inhabit, and then begin to study the observable consequences of these changes over time.

Space and Spacetime

	Space	Spacetime
Physical meaning of ds^2	$\sqrt{ds^2}$ is the distance, as would be measured by a ruler, between two infinitesimally separated points.	If $ds^2 < 0$ then $\sqrt{-ds^2/c}$ is the time elapsed on a clock that travels between two infinitesimally separated points. If $ds^2 > 0$ then $\sqrt{ds^2}$ is distance between the two points as measured by a ruler that is at rest in a reference frame in which the two points are simultaneous (have same value of the time coordinate).
Example coordinate systems and rules for ds^2 in the simplest spaces/spacetimes. For space this is Euclidean, for spacetime this is Minkowski (the spacetime of special relativity). These rules convert the differences in coordinate values between an infinitesimally separated pair of points into ds^2.	2D Cartesian: $ds^2 = dx^2 + dy^2$ 3D Cartesian: $ds^2 = dx^2 + dy^2 + dz^2$ 2D Polar: $x = r\cos\phi \;\; ; y = r\sin\phi$ $ds^2 = dr^2 + r^2 d\phi^2$ 3D Spherical: $x = r\sin\theta\cos\phi\,; y = r\sin\theta\sin\phi\,; z = r\cos\theta$ $ds^2 = dr^2 + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right)$	1+1 D: $ds^2 = -c^2dt^2 + dx^2$ 2+1D Cartesian: $ds^2 = -c^2dt^2 + dx^2 + dy^2$ 2+1D polar: $x = r\cos\phi \;\; ; y = r\sin\phi$ $ds^2 = -c^2dt^2 + dr^2 + r^2 d\phi^2$ 3+1D spherical: $x = r\sin\theta\cos\phi\,; y = r\sin\theta\sin\phi\,; z = r\cos\theta$ $ds^2 = -c^2dt^2 + dr^2 + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right)$
Examples of coordinate systems and rules for non-Euclidean spaces and non-Minkowski spacetimes	<u>Homogeneous and isotropic</u> 2D: $ds^2 = \frac{dr^2}{1-kr^2} + r^2 d\phi^2$ 3D: $ds^2 = \frac{dr^2}{1-kr^2} + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right)$ <u>Spatial part of Schwarzschild</u> 3D: $ds^2 = dr^2/(1 - r_s/r) + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right)$	<u>Homogeneous, isotropic, and *static*</u> 2+1D: $ds^2 = -c^2dt^2 + \frac{dr^2}{1-kr^2} + r^2 d\phi^2$ 3+1D: $ds^2 = -c^2dt^2 + \frac{dr^2}{1-kr^2} + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right)$ <u>Schwarzschild</u> 3+1 D: $ds^2 = -c^2dt^2(1 - r_s/r) + dr^2/(1 - r_s/r) + r^2\left(d\theta^2 + \sin^2\theta d\phi^2\right)$

HOMEWORK Problems

Problem 1.35.1

Show, by solving for x and t that the inverse Lorentz transformation is the same as the forward transformation but with $v \to -v$. Explain what this has to do with the principle of relativity.

Problem 1.35.2

Show that for straight paths in spacetime, that $(\Delta s)^2 = -c^2(\Delta t)^2 + (\Delta x)^2$ follows from $ds^2 = -c^2dt^2 + dx^2$. Hint: all straight paths in spacetime (at least the flat spacetime of special relativity we are studying now) can be parametrized via: $t - t_0 = \lambda, x = x_0 + v\lambda$.

Problem 1.35.3

Events A and B occur 10 meters and 100 ns apart in time in frame 1. If they occur 95 ns apart in frame 2, what must their spatial separation be in frame 2?

Problem 1.35.4

An astronaut leaves Earth and then returns to find their twin is much older. Assume one twin stays at rest on the Earth while the other departs at speed v and then turns around and comes back to their twin once again at speed v. Assume the time elapsed for the stay-at-home twin, between departure and return, is $t_2 - t_1$. How much time elapses for the astronaut twin between

departure and return? Draw a spacetime diagram in a frame that has the stay-at-home twin at a fixed location and make use of the invariant distance.

This page titled 1.35: Einstein Relativity is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

1.36: The Simplest Expanding Spacetime

In this chapter we begin our exploration of physics in an expanding spacetime. We begin with a reminder about how coordinates themselves are meaningless, and that physical meaning comes from the expression for the invariant distance. We start with a spacetime with just one spatial dimension that is not expanding: a 1+1-dimensional Minkowski spacetime. I expect you are familiar with such a spacetime from your prior study of special relativity, and from the previous chapter. We then generalize it slightly to describe a spacetime with one spatial dimension that is *expanding*. With additional assumptions we then calculate the age of this spacetime as well as the "past horizon."

We can label spacetimes with coordinates; for example, we could label every point in a 1+1-dimensional space with a t value and an x value. These coordinates are just labels, with no physical meaning, until we also say something about the "invariant distance" between infinitesimally separated pairs of points. For example, in a 1+1-dimensional Minkowski space with which you are familiar, it is possible to do this labeling such that the square of the invariant distance between t, x and $t + dt, x + dx$ is given by:

$$ds^2 = -c^2 dt^2 + dx^2. \tag{1.36.1}$$

Note

The physical interpretation of ds^2 is as follows:

1. For time-like separations ($ds^2 < 0$), the time elapsed on a clock that freely falls (travels with no acceleration) between the two space-time points is $\sqrt{-ds^2/c^2}$; and
2. For space-like separations ($ds^2 > 0$), the length of a ruler with an end on each of the two space-time points, at rest in the frame in which the two events are simultaneous, is $\sqrt{ds^2}$.

Box 1.36.1

Exercise 5.1.1: For the spacetime specified by Equation 1.36.1. On a plot of x vs. t (what we call a spacetime diagram) draw the trajectory of a particle that is not moving, one that is moving slowly, and then of one that is moving at the speed of light. Place the x-coordinate on the horizontal axis, as is the usual convention.

The invariant distance rule above (Equation 1.36.1) is for a *static spacetime*. Our universe is expanding. We can make a simple alteration of the invariant distance equation to describe an expanding universe:

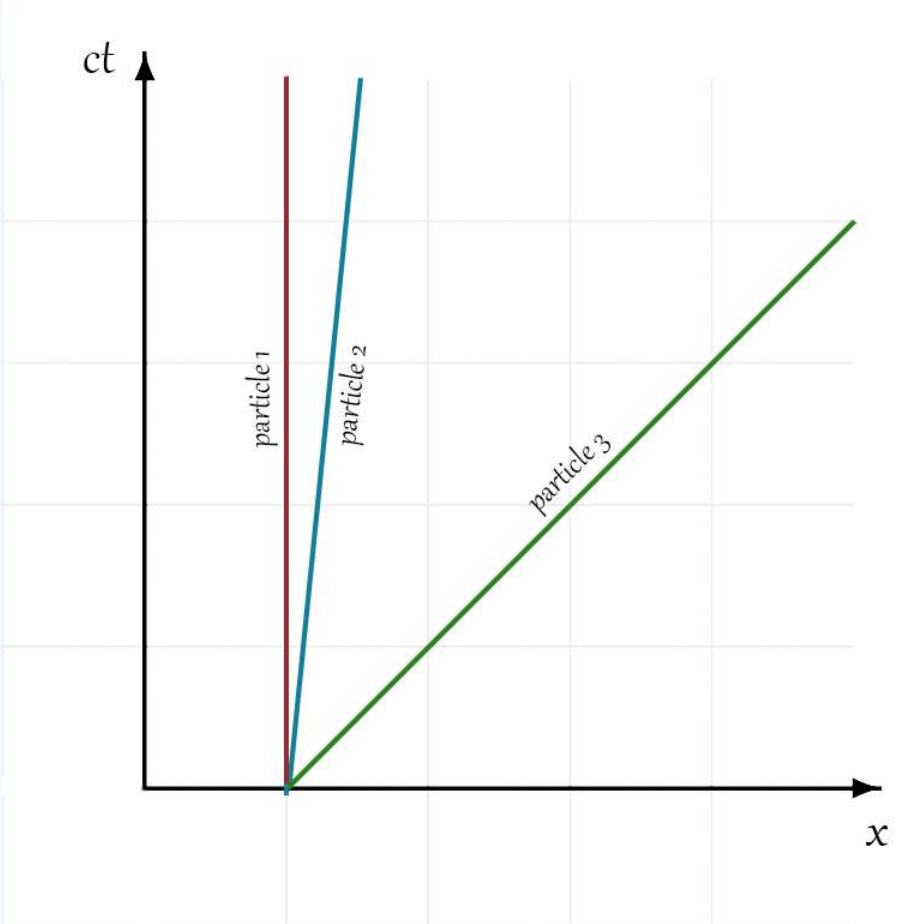
$$ds^2 = -c^2 dt^2 + a^2(t) dx^2 \tag{1.36.2}$$

with $a(t)$ a function of time. If $\dot{a} > 0$ the universe is expanding. If $\dot{a} < 0$ it is contracting. We call $a(t)$ the "scale factor."

Box 1.36.2

Exercise 5.2.1: Imagine a very small ruler instantaneously at rest in the x, t coordinate system of Equation 1.36.2 at time $t = t_1$, with one end at location $x = x_1$ and its other end at $x = x_1 + dx_1$. How long is the ruler?

Answer



We cheated a bit here and made a x vs. ct plot so that a particle moving at the speed of light has a slope of 1.

Box 1.36.3

Exercise 5.3.1: How much time elapses on a clock on a trajectory of constant x, from $t = t_1$ to $t = t_2$ for a spacetime and coordinate system with invariant distances given by Equation 1.36.2?

Answer

"at time $t = t_1$" so $dt = 0$, so $ds = a(t)dx$. Since the ruler is at rest in the given coordinate system its length is indeed given by ds at time t_1. Therefore the length of the ruler is $ds = a(t_1)dx_1$.

Box 1.36.4

Exercise 5.4.1: Still assuming Equation 1.36.2, draw the paths through spacetime of a pair of particles that are separated from each other and that are not "moving" -- that is, their x coordinate values are not changing over time. Assume $a(t)$ is an increasing function of time. What do you notice about the distance between them and how it evolves over time? Be careful not to confuse "distance between them" with the difference in the values of their spatial coordinates.

Answer

"constant x" so $dx = 0$, and then $ds = cdt$. Therefore the time elapsed on the clock is

$$\int \frac{1}{c}\sqrt{-ds^2} = \int_{t_1}^{t_2} dt = t_2 - t_1.$$

Exercise 5.4.2: Now, add in the trajectory of a light ray passing from one of these particles to the other. While sketching it out, remember that $a(t)dx$ is the distance traversed (as measured by an observer at rest in the x, t coordinate system) as the time coordinate changes by dt, which is the time elapsed as measured by an observer at rest in the x, t coordinate system. In this x vs. t diagram, does light travel in a straight line?

You should have seen in the box above that light does not travel on a straight line in this expanding spacetime as labeled with the t, x coordinates. This is kind of annoying. Often one can choose better coordinates to describe a problem in a simpler manner. For example, for a problem with spherical symmetry one can switch from Cartesian coordinates to spherical coordinates. We will do a similar thing here, introducing a coordinate called "conformal time."

The conformal time, τ, is defined via $d\tau = dt/a(t)$. In a conformal time diagram, for the expanding spacetime with which we have been working, light trajectories are straight lines. We will find that this is a very useful property.

Box 1.36.5

Exercise 5.5.1: Given a spacetime described by Equation 1.36.2, work out the invariant distance specified for τ, x labeling instead of t, x labeling. You should find $ds^2 = a^2(\tau)[-c^2 d\tau^2 + dx^2]$ where by $a(\tau)$ we just mean $a(t(\tau)$.

Answer

No solution available yet

Note that, for an observer at rest in the given coordinate system, and given our physical interpretation of the invariant distance, the equation for the invariant distance can always be written schematically as

$$ds^2 = -c^2(\text{infinitesimal time elapsed})^2 + (\text{infinitesimal spatial distance traversed})^2$$

where by "infinitesimal time elapsed" we mean as measured by a clock that is not moving in the given coordinate system and by "infinitesimal spatial distance traveled" we mean as measured by a ruler that is not moving in the given coordinate system. All observers see light traveling at the speed of light so for the path of a photon we have (infinitesimal spatial distance traversed) = c $\times$(infinitesimal time elapsed). Putting this together we can conclude that light rays travel on trajectories with $ds^2 = 0$.

Box 1.36.6

Exercise 5.6.1: Draw how light rays move on a plot of x vs. τ. Start from $ds^2 = 0$ to find the relationship between $d\tau$ and dx, then draw a trajectory consistent with that relationship.

Answer

Substituting in $dt = a(t)d\eta$ to Equation 5.2 and factoring out $a^2(t)$ gives us

$$ds^2 = -c^2 a^2(t)[d\eta^2 + dx^2].$$

Assuming a one-to-one correspondence between t and η (which one would have in an expanding universe given definition of $d\eta$) we can use $a(\eta) \equiv a(t(\eta))$ in its place and write

$$ds^2 = a^2(\eta)[-c^2 d\eta^2 + dx^2]$$

An interesting question to ask about an expanding spacetime is whether the universe ever had, in the past, the scale factor equal to zero. As this would render everything currently in the observable universe all with zero separation between them, this would be quite an extreme situation. Just to get some practice working things out in an expanding spacetime, practice that will be useful later, let's assume $\dot{a} = \kappa/a$ for κ some positive constant and see if such a universe ever had $a = 0$. Let us call the time since $a = 0$, Δt. We can then write

$$\Delta t = \int dt = \int_0^{a(t)} da/\dot{a} = \int_0^{a(t)} da(a/\kappa) = a^2(t)/(2\kappa). \tag{1.36.3}$$

Since the integral converged, we find that with the assumption given, namely $\dot{a} \propto 1/a$, the answer is yes, a finite time in the past the scale factor had the value 0. This is the singularity of the big bang. In such spacetimes we usually choose to call the zero point of

time ($t=0$), the time when $a=0$. [Note that this Δt is the time that would elapse on a stationary clock; i.e., a clock with a fixed spatial coordinate.]

Also note that we made progress with this calculation by replacing dt with $dt=da/\dot{a}$. This is a trick we will use many times to calculate a variety of things.

Another question we can ask is, "how far has light traveled since the beginning." It's interesting because nothing travels faster than the speed of light, so this tells us what the maximum distance is that any signal can propagate. We call this distance the "past horizon." Let's once again assume, for definiteness, $\dot{a}=\kappa/a$ and calculate how far light can travel. We know that for light $ds^2=0$ so we have $c^2dt^2=a^2(t)dx^2$ and therefore $cdt/a(t)=dx$ so we can write

$$\Delta x=\int dx=\int cdt/a=c\int_0^{a(t)}da/(a\dot{a})=\frac{c}{\kappa}\int_0^{a(t)}da=\frac{c}{\kappa}a(t) \tag{1.36.4}$$

(where you'll note we used the same trick again to convert an integral over time to an integral over the scale factor). Therefore we know the coordinate distance that light has traveled, Δx. That coordinate distance corresponds to a physical distance, at time t, of $a(t)\Delta x=\frac{c}{\kappa}a^2(t)$.

HOMEWORK Problems

Problem 1.36.1

Derive the phenomenon of Lorentz contraction using the invariance of the invariant distance. [Do not assume an expanding universe; assume $ds^2=-c^2dt^2+dx^2$]. The trick to doing this is careful choice of the two events (points in spacetime) for which to calculate their invariant distance. Imagine a ruler moving with respect to an observer at speed v, with the ruler oriented so that it is parallel to the relative velocity. Take event 1 to be when/where the front end of the ruler is at the same spacetime location as the observer, and event 2 to be when/where the back end of the ruler is at the same spacetime location as the observer. By calculating the invariant distance in the observer's rest frame and the ruler's rest frame you should find that the length of the ruler as determined by the observer is $L'=L/\gamma$ where L is the length of the ruler in its rest frame.

Problem 1.36.2

Assume that the scale factore evolves via $\dot{a}=\kappa a$ for κ a positive constant. (Note that this is a *different* assumption than the previous $\dot{a}=\kappa/a$). Show that in this spacetime the universe never has $a=0$. Do so by showing that the amount of time between $a=0$ and any finite a is infinite; i.e., show that the appropriate definite integral does not converge.

Problem 1.36.3

Assume $ds^2=-c^2dt^2+a^2(t)dx^2$ and once again that $\dot{a}=\kappa a$ for κ a positive constant. Our universe appears to be moving asymptotically toward such a case (although except with a 3-dimensional space instead of a 1-dimensional space). Determine what we call the "future horizon." If a light signal is sent out at time t_1 from x_1, in the positive x direction, to what value of x_2 will it get given an infinite amount of time? The distance between x_1 and x_2 at time t_1, $a(t_1)(x_2-x_1)$, is called the future horizon.

This page titled 1.36: The Simplest Expanding Spacetime is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

- **Current page** by Lloyd Knox is licensed CC BY 4.0.
- **2.5: S05. Spacetime - SOLUTIONS** by Lloyd Knox is licensed CC BY 4.0.

1.37: Redshifts

We introduced expanding, non-Euclidean spacetimes in the previous chapter. Now we begin to work out observational consequences of living in such a spacetime. In this and the next two chapters we will derive Hubble's Law, $v = H_0 d$. In fact, we will derive a more general version of it valid for arbitrarily large distances.

In Minkowski space, the invariant distance, ds, between spacetime point (t, x, y, z) and another one at $(t + dt, x + dx, y + dy, z + dz)$ is given by:

$$ds^2 = -c^2 dt^2 + dx^2 + dy^2 + dz^2 \tag{1.37.1}$$

We call it invariant because it is invariant under Lorentz transformations. Physically, for $ds^2 > 0$, ds is the length of a ruler with one end on each of the two space-time points, if the ruler is at rest in a frame in which the two space-time points are simultaneous. For $ds^2 < 0$, $\sqrt{-ds^2}/c$ is the amount of time that elapses on a clock that passes from one space-time point to the other.

In spherical coordinates the above expression for the invariant distance becomes:

$$ds^2 = -c^2 dt^2 + dr^2 + r^2 \left(d\theta^2 + \sin^2\theta d\phi^2\right) \tag{1.37.2}$$

More generally, the invariant distance in a spatially homogeneous and isotropic Universe can be written as:

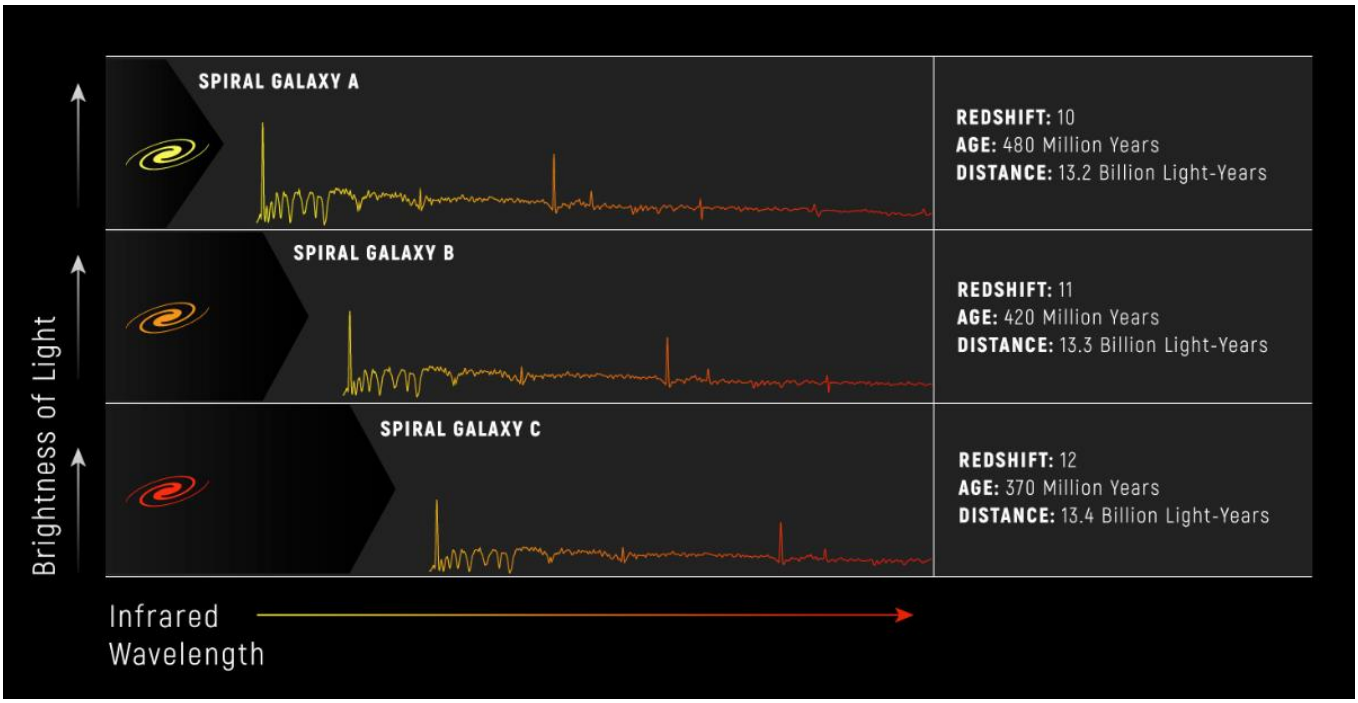
$$ds^2 = -c^2 dt^2 + a^2(t) \left[\frac{dr^2}{1 - kr^2} + r^2 \left(d\theta^2 + \sin^2\theta d\phi^2\right)\right] \tag{1.37.3}$$

Such spacetimes are known as Friedmann-Robertson-Walker (FRW) models, or sometimes just Robertson-Walker, and sometimes Friedmann-Robertson-Walker-Lemaitre models.

We are now ready to define "redshift" and derive the relationship between it and the expansion of space. Redshift is a quantification of the stretching of the wavelength of light. We use the symbol z to denote redshift and define it as:

$$z \equiv (\lambda_{\text{received}} - \lambda_{\text{emitted}})/\lambda_{\text{emitted}}. \tag{1.37.4}$$

We see in the figure below (from NASA and Space Telescope Science Institute) some model spectra of galaxies with redshifts of 10, 11, and 12. One can see in the image that the same spectral features appear at longer wavelengths as the redshift increases. These features all had the same wavelength when emitted, but those wavelengths get stretched by the expansion to longer wavelengths. The further away the objects, the greater the stretching that occurs. We will explore this relationship between distance and redshift in chapter 8.



Next we will derive the relationship between redshift and expansion for light that leaves an object at rest at time t_e and is observed today by an observer at rest at time t_r. We'll get there with the following two exercises.

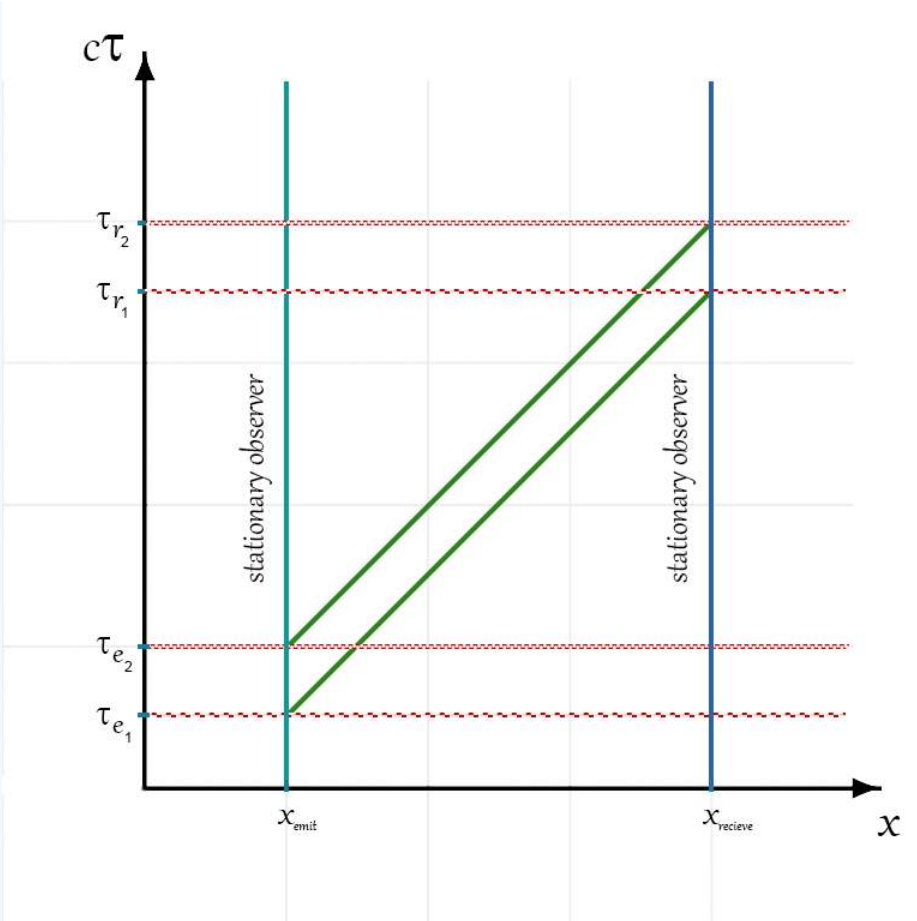
Note

By "at rest" we mean at rest in the coordinate system that has an invariant distance of the form in Equation 1.37.3. This frame is called "cosmic rest." Note that for a frame that is moving with respect to cosmic rest, slices of constant time would be different, and we'd no longer have this simple form where the scale factor only depends on the time coordinate.

Box 1.37.1

Exercise 6.1.1: Show that if Δt_e is the time interval between emitted light pulses (as measured by a stationary observer located where the emission is happening) and if Δt_r is the time interval between reception of first pulse and second pulse (as measured by a stationary observer located where the reception is occurring) then $\Delta t_r/\Delta t_e = a(t_r)/a(t_e)$. To do so, use conformal time defined by $d\tau = dt/a(t)$ and draw the pulse trajectories on an x vs. τ diagram. By stationary observer we mean one that is at a constant value of x; i.e., is at rest in the cosmic rest frame. You can assume that Δt_r and Δt_e are very short time scales compared to the time scale over which $a(t)$ changes appreciably. Practical applications of this result often have $a(t)$ changing on billion-year time scales and the Δts shorter than nanoseconds so this assumption is well justified!

Answer



It's clear from the graphic that $\Delta\tau_e = \Delta\tau_r$. We can integrate up $d\tau = dt/a(t)$ by approximating $a(t)$ as constant over these short time intervals to get $\Delta\tau = \Delta t/a(t)$. Then we can easily see that $\Delta t_r/\Delta t_e = a(t_r)/a(t_e)$.

Box 1.37.2

Exercise 6.2.1: Imagine propagation of electromagnetic waves. Use the above result to show that the wavelength of the waves emitted at time t_r and observed at time t_e is stretched so that:

$$z \equiv (\lambda_{\text{received}} - \lambda_{\text{emitted}})/\lambda_{\text{emitted}} = a(t_r)/a(t_e) - 1$$

Hint: think about what happens to the period of the wave first, and then go from that to wavelength using the fact that wavelength is proportional to period.

Answer

Imagine successive crests of a single wave. Map one crest, and then the subsequent one, separated in time by the period of the emitted wave T_e, on to the two pulses you considered in the previous exercise. The time between the pulses upon emission is $\Delta t_e = T_e$. The time between the reception of the first pulse and reception of the second pulse is the period of the wave upon reception $T_r = \Delta t_r = (a(t_r)/a(t_e))\Delta t_e = (a(t_r)/a(t_e))T_e$. Wavelength is proportional to period so we also have $\lambda_r = (a(t_r)/a(t_e))\lambda_e$ or

$$\frac{\lambda_r}{\lambda_e} = \frac{a(t_r)}{a(t_e)}.$$

We already know that $z \equiv (\lambda_{\text{received}} - \lambda_{\text{emitted}})/\lambda_{\text{emitted}}$, so we can rewrite it as

$$z \equiv \frac{\lambda_r}{\lambda_e} - 1 = \frac{a(t_r)}{a(t_e)} - 1$$

Box 1.37.3

Exercise 6.3.1: The most distant object for which a redshift has been measured is called a gamma-ray burst and it has a redshift of $z = 8.2$. By what factor has the universe expanded since light left that object?

Answer

The universe has expanded by a factor

$$a(t_r)/a(t_e) = 1 + z = 1 + 8.2 = 9.2.$$

We have just worked out the amazing fact that if we can identify a spectral line and measure its wavelength, then we can directly determine how much the universe has expanded since light left the object. Rearranging the result from Exercise 6.2.1 a little we can express this amazing fact as an equation:

$$1 + z = \lambda_r/\lambda_e = a(t_r)/a(t_e). \tag{1.37.5}$$

Look again at the figure above and notice that the more distant the object, the higher the redshift. This is because the more distant the object, the more time it took light to get to us from the object, and therefore the smaller the scale factor was at the time the light left the object, t_e.

Before closing this section, we remind the reader of how redshifts due to the Dopper effect are related to speed -- a result we will use later in deriving Hubble's Law. If a source is moving away from you at speed v and emitting pulses with period T, then the second pulse has to travel a distance vT further to get to you than was the case for the first pulse. So it's arrival will be delayed by a time vT/c. Thus the period for the arriving pulses is $T(1 + v/c)$. Since wavelength is proportional to period this means the wavelength is stretched by a factor of $1 + v/c$, which means, by definition of the redshift z, that $z = v/c$. Note that our derivation has ignored the effect of relativistic time dilation. If the source had period T at rest, then if it were moving with speed v with respect to us, in our frame the period would be stretched to γT and so the complete expression for z from the Doppler effect is $1 + z = \gamma(1 + v/c)$. But we are only interested in this expression for small v/c, for which $\gamma \sim 1$.

Summary

1. In an expanding homogeneous and isotropic universe, the ratio between the wavelength of light emitted by an observer at cosmic rest, λ_e at time t_e and the wavelength as measured by an observer at cosmic rest λ_r at time t_r is given by

$$1 + z \equiv \lambda_r/\lambda_e = a(t_r)/a(t_e) \tag{1.37.6}$$

   where $a(t)$ is the scale factor at time t and the above equation defines the redshift z. We can think of this as the wavelength stretching with the expansion.
2. In a non-expanding universe, a source moving away from an observer with $v/c \ll 1$ has its light redshifted (wavelength stretched) by

$$\lambda_r/\lambda_e = 1 + v/c \tag{1.37.7}$$

   which is the normal Doppler effect you have studied before. What this implies is that if we interpret a small redshift z caused by the expansion of space as due to an ordinary motion-induced Dopper effect we will set $z = v/c$. We will use this relationship later in our derivation of Hubble's Law: $v = H_0 d$.

Additional Resources

You can look at images and spectra of galaxies in the Sloan Digital Sky Survey here. The spectra include identification of emission and absorption lines (with the atoms or ions responsible for them) and measurements of the redshift. To find the spectra and images, look for the table and click on the Object ID.

Quasar absorption line systems have particularly interesting spectra. You can read about them here.

This page titled 1.37: Redshifts is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Lloyd Knox.

- **Current page** by Lloyd Knox is licensed CC BY 4.0.

- **2.6: S06. Redshifts - SOLUTIONS** by Lloyd Knox is licensed CC BY 4.0.

Index

B

Bayes' Theorem
- 1.17: Cosmological Data Analysis

Big Bang
- 1.18: The Early Universe
- 1.21: Hot and Cold Relics of the Big Bang
- 1.23: Big Bang Nucleosynthesis - Predictions

boosts (relativity)
- 1.34: Galilean Relativity

C

Cepheid variables
- 1.16: Parallax, Cepheid Variables, Supernovae, and Distance Measurement

chapters
- 1.2: Spacetime Geometry
- 1.3: Redshifts
- 1.4: Spatially Homogeneous and Isotropic Spacetimes
- 1.5: Euclidean Geometry
- 1.6: Distances as Determined by Standard Candles
- 1.7: The Distance-Redshift Relation
- 1.8: Dynamics of the Expansion
- 1.9: A Newtonian Homogeneous Expanding Universe
- 1.10: The Friedmann Equation
- 1.11: Particle Kinematics in an Expanding Universe - Newtonian Analysis
- 1.13: Energy and Momentum Conservation
- 1.14: Pressure and Energy Density Evolution
- 1.18: The Early Universe
- 1.19: Equilibrium Statistical Mechanics
- 1.20: Equilibrium Particle Abundances
- 1.21: Hot and Cold Relics of the Big Bang
- 1.32: Euclidean Geometry
- 1.33: Curvature
- 1.35: Einstein Relativity
- 1.36: The Simplest Expanding Spacetime
- 1.37: Redshifts

CMB
- 1.26: The Spectrum of the CMB

CMB Power Spectrum
- 1.27: Cosmic Microwave Background Anisotropies

Comoving angular diameter distance
- 1.15: Distance and Magnitude

Comoving distance
- 1.15: Distance and Magnitude

Coordinate distance
- 1.15: Distance and Magnitude

COsmic Background Explorer
- 1.26: The Spectrum of the CMB

Cosmic Inflation
- 1.25: Introduction to the Cosmic Microwave Background

cosmic microwave background
- 1.25: Introduction to the Cosmic Microwave Background

curvature
- 1.33: Curvature

D

dark energy
- 1.12: The Evolution of Mass-Energy Density and a First Glance at the Contents of the Cosmos

E

Euclidean Geometry
- 1.5: Euclidean Geometry
- 1.32: Euclidean Geometry

F

Far Infrared Absolute Spectrometer
- 1.26: The Spectrum of the CMB

FIRAS
- 1.26: The Spectrum of the CMB

Friedmann equation
- 1.10: The Friedmann Equation

G

Galilean Boost
- 1.34: Galilean Relativity

Galilean relativity
- 1.34: Galilean Relativity

H

homogeneity
- 1.9: A Newtonian Homogeneous Expanding Universe

Hubble's Law
- 1.3: Redshifts
- 1.4: Spatially Homogeneous and Isotropic Spacetimes
- 1.6: Distances as Determined by Standard Candles
- 1.9: A Newtonian Homogeneous Expanding Universe
- 1.37: Redshifts

I

Incomplete
- 1.2: Spacetime Geometry
- 1.3: Redshifts
- 1.4: Spatially Homogeneous and Isotropic Spacetimes
- 1.16: Parallax, Cepheid Variables, Supernovae, and Distance Measurement
- 1.36: The Simplest Expanding Spacetime
- 1.37: Redshifts

L

Lorentz boosts
- 1.35: Einstein Relativity

luminosity
- 1.15: Distance and Magnitude

luminosity distance
- 1.6: Distances as Determined by Standard Candles
- 1.15: Distance and Magnitude

N

nucleosynthesis
- 1.23: Big Bang Nucleosynthesis - Predictions

P

parallax
- 1.16: Parallax, Cepheid Variables, Supernovae, and Distance Measurement

parallel postulate
- 1.33: Curvature

past horizon
- 1.2: Spacetime Geometry
- 1.36: The Simplest Expanding Spacetime

Physical distance
- 1.15: Distance and Magnitude

R

Redshifts
- 1.3: Redshifts
- 1.4: Spatially Homogeneous and Isotropic Spacetimes
- 1.37: Redshifts

S

standard candles
- 1.6: Distances as Determined by Standard Candles

supernovae
- 1.16: Parallax, Cepheid Variables, Supernovae, and Distance Measurement

T

The Friedmann Equation
- 1.10: The Friedmann Equation

W

WIMPS
- 1.21: Hot and Cold Relics of the Big Bang

Glossary

absolute magnitude | The apparent magnitude an object would have if it were at a distance of 10 parsecs. cf magnitude.

absolute zero | An idealized temperature at which there is no energy left in a given system. 0 Kelvin is absolute zero, which is -273.15 ^{o}C, or -459.67 ^{o}F. Modern quantum physics precludes real systems from reaching absolute zero.

absorption | The process by which light or other electromagnetic radiation is absorbed by an atom, giving its energy to the atom in the process.

absorption line spectrum | A spectrum showing dark lines in some narrow color regions (wavelengths). The lines are formed when atoms absorb the light at specific wavelengths.

accelerate | A change in uniform motion, either from slowing down or speeding up, or changing direction.

accretion disk | A disk of matter that forms when material is transferred to a small gravitating body, such as a black hole or protostar. For black holes, the disks form outside the event horizon. For other objects, such as neutron stars or protostars, the disks can extend down to the stellar surfaces. Friction (collisions) within the disks heat them and allow material to flow inward while angular momentum flows outward. Accretion disks often emit a wide range of different types of electromagnetic radiation including infrared, UV, and x-rays.

accuracy | In science, the closeness of a measurement of some quantity to its true value. This differs from precision.

active galaxy | A galaxy with a very bright, energetic nucleus. The evidence suggests that they are powered by the release of gravitational energy as material falls onto a central black hole. The range in mass from several million to several billion times the mass of the Sun. Sometimes active galaxies are called AGN, for active galactic nuclei.

alpha radiation | a type of radiation emitted during radioactive decay, an alpha particle is a ^{4}He nucleus

anti-particle | The antimatter complement to a particle, mostly having identical properties to the particle, but with opposite electric charge.

apparent magnitude | How bright an object appears as seen by an observer. cf magnitude.

arcminute | A measure of angular size based on a circle. A full circle has 360°, which can be divided into 60 equal parts, each part being 1 degree. An arcminute, in turn, is one sixtieth (1/60) of a degree.

arcsecond | One sixtieth (1/60) of an arcminute.

asteroid | A rocky object in space that can be a few meters to hundreds of kilometers wide.

astronomical unit | The average distance between Earth and the Sun, equal to 149,597,870.7 kilometers. Abbreviated AU.

atom | A basic physical building block of matter in the Universe, which is composed of electrons, protons, and neutrons.

atomic number | Indicates the number of protons in the nucleus of an atom. The atomic number defines a chemical element.

AU | See Astronomical Unit.

background counts | Sources of light detected by a CCD that are extraneous to the celestial object being studied.

band-pass | A specific range of electromagnetic frequencies, often used to describe a satellite's viewing capabilities.

BBN | See Big Bang Nucleosynthesis.

beta radiation | a type of radiation emitted during radioactive decay, an electron or a positron is emitted

Big Bang Nucleosynthesis (BBN) | The processes in the early Universe that created the lightest chemical elements. Also, the period of time over which these processes were active.

Big Bang Theory | The theory that suggests that the Universe at early times was compressed into an extremely hot, dense state. Then, for reasons currently unknown, the universe expanded and cooled to its present condition, and it continues to expand and cool at the present time.

big chill | An end-of-Universe scenario where galaxies continue to move away from each other and the temperature of the universe continues to cool.

big crunch | An end-of-Universe scenario where gravitational attraction causes galaxies to eventually start moving toward each other.

big rip | An end-of-Universe scenario where dark energy becomes so strong that the expansion rips apart galaxies, even at the atomic level.

binary | The word binary simply means there are two of something. When applied to a star system, it means that two stars orbit their common center of mass.

binary star system | A system of two stars, very close to each other, that orbit around their common center of mass. Such systems are quite common.

binding energy | The energy required to break up an atomic nucleus into its constituent protons and neutrons.

black hole | A region of space within which the force of gravity (space-time curvature) is so strong that nothing, not even light, can escape from it.

blackbody | A theoretical object that is a perfect absorber and emitter of electromagnetic radiation. Such an object would emit so-called blackbody radiation.

blackbody radiation | Radiation produced by a blackbody. The intensity at each wavelength follows a distribution that depends on the temperature of the object.

blackbody spectrum | The spectrum of the radiation emitted by a blackbody depends upon the temperature of the black body. cf. blackbody radiation.

blueshift | An apparent shift of spectral lines in the radiation emitted by an object toward shorter wavelengths. Often caused by motion of the object toward the observer, or vice versa. See also Doppler effect.

brightness | The amount of light an observer sees emitted from an object like a star. Brightness is measured in watts per square meter (W/m^{2}).

brown dwarf | A cosmic object that is too small to be a star and too large to be a planet. Brown dwarfs have the same composition as stars, but because of their low mass are unable to sustain nuclear fusion at their cores (assuming that they ever manage to get fusion started).

bulge | The central region of a spiral galaxy.

causality | the idea that some events are the direct cause of other events, and that the cause must precede its effect.

CCD | See charge-coupled device.

chameleon particles | A hypothetical particle which is postulated as a dark energy candidate. The particle is chameleon-like in that its mass can change.

charge-coupled device (CCD) | An instrument that can act as a camera sensor. It works by converting the light that falls onto it into electrical signals in individual picture elements (pixels), which are generally arranged in a square array.

chronometer | A watch that has been specifically designed to keep very accurate time.

coefficient | A number that multiplies another number or expression.

coma | The cloud-like ball of gas and dust surrounding a comet's nucleus.

comet | Dusty bodies of ice that orbit a star. We typically imagine comets with their characteristic tails, but the tails only form when comets approach close to a star. Comets have three distinct tails: the dust tail is composed of dust pushed out by radiation pressure from the star, the ion tail is composed of particles evaporated by solar winds and pushed back, and a tail of sodium escapes from the dust. The sodium tail is not visible to the naked eye. These tails point in slightly different directions but always away from the star.

concordance model | Our currently accepted and most commonly used cosmological model.

conservation of energy | a physical law that says that energy cannot be created or destroyed. Because of this law, any naturally occurring physical process can only transfer energy from one part of a system to another part; the total energy must remain the same.

constellation | A region in the sky that has been officially defined by the International Astronomical Union. Constellations are used by astronomers to designate a region in the sky, similar to the way countries are used to designate regions on Earth. Most constellations, especially in the Northern Hemisphere, have historical origins related to the myths of Ancient Greece.

continuous spectrum | A spectrum that is an unbroken range of wavelengths. An object emitting some light in every wavelength, such as a blackbody, produces a continuous spectrum.

continuum | See continuous spectrum.

convergence point | The point on the horizon where parallel lines seem to converge. In astronomy, it is the point to which an extended object such as a star cluster seems to converge as it moves away from the observer. If the motion is toward the observer, then the object seems to expand away from the convergence point.

cosmic distance ladder | A hierarchal process used by astronomers to determine the distances to very distant astronomical objects based on known distances of similar objects that are closer.

Cosmic Microwave Background (CMB) | This is the radiation left over from the big bang. It was produced early in the age of the universe, when the average density and temperature were much higher than today. The expansion of the universe has cooled the radiation to its current temperature of about 2.7 Kelvin.

cosmic rays | high-energy charged particles from outer space. These include protons, neutrons, atomic nuclei, and subatomic particles.

cosmology | The astrophysical study of the history, structure, and evolution of the universe.

critical density | A special value of density that causes the Universe to have zero curvature.

dark current | The small but detectable charge that accumulates in a CCD, even when it is not exposed to light. This is a source of noise in astronomical instruments that must be removed.

dark energy | a hypothetical form of energy or property of space that causes the expansion rate of the universe to accelerate.

dark matter | Making up approximately 80% of the matter in the Universe, dark matter does not appear to radiate or absorb any light and is only detectable indirectly by the effect its gravity has on visible objects.

dark nebula | cold, relatively dense clouds of interstellar gas and dust

data | The factual information a scientist collects that is related to the hypothesis he or she is testing. Data can be direct measurements of properties, for example, the height of plants one month after seeds have been planted. Data can also be observations of patterns, for example, the behavior of animals when they encounter a predator. To determine whether a particular measurement or observation is a rule rather than an exception, a scientist will often repeat measurements or observations to gather additional data.

degeneracy | A combination of values having the same properties.

density | A measure of how much matter is packed into a given volume. High-density objects have more material packed into a given volume than low-density objects.

detector | A device, or devices, used to capture photons, or in some cases, other particles like protons, electrons, etc.

diffraction grating | Parallel slits or grooves etched on an optical surface that cause light to bend and create an interference pattern after passing through the grating. This spreads the light out into a rainbow of colors.

disk | A thin, roughly circular plane in which the majority of stars, gas, and dust of a spiral galaxy are contained.

distance | The length of space between two objects.

Doppler effect | The Doppler effect is the apparent change in the wavelength and frequency of sound or light depending on whether the source is moving toward or away from you. The faster you or the source is moving, the more profound the impact. You may have experienced an example of this with sound waves if you have ever heard a higher pitch sound from a car moving toward you, and a lower pitch noise when it is moving away.

Doppler shift | The shift in the frequency or wavelength of a wave due to the Doppler effect.

dust | mixture of molecules, such as silicates, graphite, iron, and other compounds

dwarf galaxy | A galaxy that contains somewhere around several billion stars, as opposed to the hundreds of billions of stars found in a large galaxy like the Milky Way.

dwarf planet | An astronomical object that orbits its parent star and that has enough mass for gravity to make it spherical in shape, but has not cleared its orbital path of debris.

ecliptic | The name given to the plane in which the Earth travels as it orbits the Sun.

electromagnetic (EM) radiation | Another term for light, including visible light and invisible forms from radio up through gamma-rays. See electromagnetic spectrum.

electromagnetic (EM) spectrum | This is the continuum of waves of light, which range from very low-frequency and low-energy radio waves to very high-frequency and high-energy gamma rays. The kind of light we are familiar with is visible light, which is a tiny sliver in the middle of the EM spectrum.

electromagnetic force | The electromagnetic (or EM) force is one of the four known universal forces, along with gravity and the strong and weak nuclear forces. The EM force holds all the molecules and cells in your body together and is the result of interactions between charged particles (protons and electrons) within the atoms and molecules.

electromagnetic waves | Another term for light. Light waves are fluctuations of electric and magnetic fields in space.

electron | A fundamental subatomic particle that is commonly found in the outer regions of an atom and is negatively charged. Electrons are a type of lepton.

electron cloud | The region around the nucleus of an atom that the electrons occupy.

electron volt | A unit of energy. 1 eV = 1.602×10^{-19} joules. It is the energy gained by an electron (or a proton) falling through an electrical potential difference of one volt. Visible light photons have energies of about one electron volt.

element | A substance that is composed of a single type of atom.

elementary particle | A particle which is not made up of any other particles or substructure, such as quarks and leptons.

elliptical galaxy | These galaxies range in shape from nearly spherical to flattened disks. They are characterized by an old population of stars and have very low rates of star formation, meaning very few stars are being born in them. Ellipticals contain little or no cool gas or dust.

emission | The production of light, or more generally, electromagnetic radiation, by an atom or other object.

emission line spectrum | A spectrum consisting of bright lines at certain wavelengths separated by dark regions in which there is no light.

emission nebula | clouds of hot interstellar gas that emit light

energy | Energy is the ability to do work. The SI unit for energy is the joule (J), where 1 J = 1 kg $(m/s)^2$ = 1 N m. The eV (see electron volt) is another common unit of energy: 6.242×10^{18} eV = 1 J. Joules are also related to watts (W), the SI unit for power, via: 1 W = 1 J/s. One exploded ton of TNT is equivalent to 4.2×10^9J, or 2.6×10^{28} eV.

entropy | a physical measure of the disorder of a system, often computed by measuring the number of possible configurations of the system.

error ellipse | A region on a plot that covers all allowable values for the specified cosmological parameters given the measurement uncertainties.

escape speed | The minimum speed required to escape the gravitational pull from a massive object.

eV | See electron volt.

event | in special relativity, an event is a location (t, x, y, z), in spacetime, that describes the position and time of an occurrence.

event horizon | The region near a black hole where the escape speed becomes the speed of light. Anything that crosses the event horizon cannot escape the gravitational pull of the black hole.

evolve | To change over time.

excited state | An energy state in a quantum system that lies above the lowest available state, or ground state. In an atom, an electron that has gained energy and moved from its lowest state into a higher energy state is said to be in an excited state. The atom must release energy, often by emitting a photon, for the electron to return to the ground state.

exponent | The value to which a base number is raised. For example, in the number 10^4, the 4 is the exponent, 10 is the base.

extra-solar planet | A planet that orbits a star other than the Sun and is therefore not within our Solar System.

field bosons | Fundamental particles of integral spin (0, +/-1, +/-2, etc) that carry force between other particles. Sometimes called *field bosons*. One example is the photon, which carries the electromagnetic force. Another is the gluon, which carries the strong nuclear force.

filter | A device that can be placed in front of a camera that lets certain wavelengths of light pass through and blocks other wavelengths.

flux | A measure of the amount of energy given off by an astronomical object per unit time per unit area. Because the energy is measured per time and area, flux measurements make it easy for astronomers to compare the relative energy output of objects with very different sizes or ages.

frame (of reference) | in physics, a frame of reference refers to the coordinate system used to conduct measurements, actual or imagined. These coordinates could be fixed to a laboratory that is stationary on Earth, or they could be on a moving object like a plane flying through the air. They could even be on some other planet, or even on an imagined spaceship that travels between the planets or stars. All measurements are only meaningful if referred to some standard of measure in some frame of reference.

frequency | A property of a wave that describes how many wave patterns, or cycles, pass by a point in a given time. Frequency is often measured in hertz (Hz), where one hertz is one cycle per second.

fusion | The process of merging multiple objects into one. In nuclear fusion, light atoms are forced together to form heavier atoms. For instance, hydrogen can be fused to form helium. Helium in turn can be fused into the heavier atom carbon, and so on. This process releases large amounts of energy and is what powers stars. However, fusion requires very high temperatures. That is why a tube of hydrogen or helium gas in a laboratory will not spontaneously undergo nuclear fusion.

galaxy | A gravitationally bound system of stars, their satellites, dust, gas, and dark matter that contains a supermassive black hole at its center. There are three general types of galaxies; spiral, elliptical, and irregular.

galaxy cluster | A group of two or more galaxies that are bound by gravity.

galaxy halo | A spherical distribution of stars, with very low density, in which galaxies are embedded. Different than dark matter halo.

gamma | The factor in special relativity by which time and space are stretched or compressed by relative motion. Also, the ratio of the energy of a particle to its rest energy. The gamma factor depends on the relative velocity between reference frames approximately equal to x for small angles x (when x is measured in radians).

gamma rays | The very highest energy end of the electromagnetic spectrum, with the shortest wavelengths. Gamma rays typically have energies a few hundred times larger than low-energy x-rays and wavelengths shorter than a few hundred picometers (pm, 10^{-12} m).

gas | atoms and small molecules, primarily hydrogen

gas giant | A planet that is at least five times as massive as Earth and comprised mostly of gases, such as hydrogen and helium.

Gedankenexperiment | German, "thought-experiment." These are illustrative scenarios used to guide one's thinking when considering physical situations, especially when thinking about relativistic systems. These were often employed by Albert Einstein when developing his physical intuition regarding problems related to relativity.

GeV | Gigaelectron volt, or 1 billion electron volts.

Giga (prefix) | A billion, 10^9. Denoted as G, e.g., GeV.

globular clusters | Exceptionally dense, spherical clusters of old stars that are gravitationally bound, located in galaxy halos.

gluon | A subatomic particle that carries the strong nuclear force between quarks.

gravitational constant | See Universal Constant of Gravitation.

gravity | The universal force of attraction between all matter.

Great Red Spot | A giant cyclone that has existed on Jupiter for at least 300 years. It is not known whether it will ever disappear.

greatest elongation | When the Sun, Earth and a planet interior to Earth's orbit are positioned such that the planet appears as far as possible from the Sun in the sky as seen from Earth. Only Venus and Mercury can attain such a configuration.

ground state | The ground state is the lowest energy state available in an atom or other quantum system.

H-R diagram | See Hertzsprung-Russell diagram.

half-life | The average time required for half of a sample of radioactive atoms to decay

halo | See dark matter halo or galaxy halo.

helium-3 | A form of helium that has 2 protons and 1 neutron. He-3, which is used in nuclear fusion research, is rare on Earth.

hertz | Abbreviated "Hz". The derived SI unit of frequency, defined as cycles per second.

Hertzsprung-Russell diagram | A plot of the brightness of stars versus their surface temperature or spectral class. Abbreviated to H-R diagram.

histogram | A graph that displays the density of data, generally the frequency of occurrence of a range of events, such as the frequency of photons of specific wavelengths striking a detector per second (example histogram below). A familiar use of a histogram is the "bell curve" used to show the distribution of grades on a test, for example.

HST Key Project | Hubble Space Telescope (HST) had several chief missions, called Key Projects. One such mission was to accurately determine the extragalactic distance scale, beginning with galaxies in the Virgo Cluster.

Hubble's Law | The relationship between the recession velocity of a galaxy and its distance from the observer. The farther away a galaxy is from an observer, the faster it is moving away. The law is named for Edwin Hubble, who first published it in 1929 (Proc. Nat. Acad. of Sci, Vol 15, March 15, 1929).

hydrostatic equilibrium | The state in a fluid, such as the gas in a star, or the atmosphere or ocean of a planet, in which the compressional force of gravity is everywhere offset by the outward force of pressure in the fluid.

Hz | See hertz.

image processing | The process by which data gathered from a CCD or other electronic imaging detector are converted into images that can be interpreted by an astronomer.

imaging | Scientific technique that results in photographic or computerized representations of data.

imprecision | Lack of precision or repeatability. cf. precision.

inertial frame | a frame of reference that moves at constant velocity. An inertial frame does not change its speed or direction.

infrared light | Abbreviated "IR". The band of the electromagnetic spectrum intermediate between optical and microwaves, with wavelengths in the micron (μm, 10^{-6}m) range. Infrared is correspondingly more energetic than microwaves, but less energetic than optical.

intensity | See flux.

International System of Units | Abbreviated SI Units, also called the metric system, it is the scientific standard of carefully defined units of measurement. The SI unit of length is the meter, of time is the second, and of mass is the kilogram. Many other units are used to measure other quantities.

invariance (invariant) | see Principle of Invariance.

Inverse Square Law | A relationship that states that the flux, or apparent brightness, of an object decreases as the inverse of the square of its distance to the observer.

ionization | The process of stripping electrons from an atom.

ionized gas | see Plasma.

irregular | A galaxy that does not fit into any of the other categories (spiral or elliptical).

isotopes | Atoms with the same number of protons but different numbers of neutrons.

jovian planets | Jupiter-like planets that do not have solid surfaces. They are composed primarily of helium and hydrogen. They typically have radii larger than 10,000 km (6,213.7 miles) with a mass over 1×10^{25} kg. These gaseous planets also have rings and many moons.

k | See kilo.

K | See Kelvin.

Kelvin | The SI unit for temperature. The temperature at which water freezes on the surface of Earth is 273.15 kelvin, and the temperature that water boils is 373.15 kelvin. Zero on the Kelvin scale is the theoretical point where all motion ceases in classical thermodynamics (absolute zero). The name honors the 19th century Scottish physicist William Thomson, who is more commonly known as Lord Kelvin.

keV | 1 kilo electron volt is equal to 1,000 electron volts, and is a unit of energy convenient for describing x-ray energies.

kg | See kilogram.

kilo (prefix) | A thousand; 10^3. Denoted as k, e.g., keV.

kilogram | Abbreviated "kg". The SI unit of mass. The kilogram is the only SI unit still maintained by a physical artifact (a platinum-iridium bar) kept in the International Bureau of Weights and Measures at Sevres, France. One kilogram is equivalent to 1,000 grams or about 2.2 pounds; the mass of a liter of water.

kilometer | A unit for measuring length, one thousand meters. Abbreviated km.

Kuiper belt | The region of the solar system past Neptune, from approximately 30 AU to 100 AU. Objects in this region are called Trans-Neptunian Objects (TNO).

length contraction | the distortion of measured lengths between inertial frames. Observers in different inertial frames will measure distances along the direction of relative motion between their frames to be longer in their own frame than in any frame moving with respect to their own.

lepton | A fundamental particle with little mass, or possibly no mass in some cases. Electrons, muons and tauons are all leptons, as are their associated neutrinos.

light speed | See Speed of Light.

light-hour | The distance that light travels in a vacuum in one hour, approximately equal to one billion kilometers (1×10^9 kilometers).

light-minute | The distance that light travels in one minute, which is approximately 18 million kilometers (1.8×10^7 km).

light-year | The distance that light travels in one year, which is about 10 trillion kilometers (9.45×10^{12} km). Light-years are a convenient unit of measure for most astronomical distances.

lookback time | The delay of time, due to the finite speed of light, required for light to travel from its source to its observer. The lookback time for terrestrial objects is negligible, but for astronomical objects, it grows with their distance, from about 8 minutes for light from the Sun to billions of years for distant galaxies.

luminosity | The amount of energy an object, like a star, radiates per unit time. This is usually measured in watts, just like a light bulb.

magnification | The process of enlarging the appearance of an object through various optics and lenses.

magnitude (astronomy) | A measure of brightness. Counterintuitively, the brighter an object is, the lower its magnitude. A first magnitude star is about 2.5 times as bright as a second magnitude star, and so on. The brightest object in the sky is, of course, the Sun, with a magnitude of-26.73. The full moon is -12.6, and with the naked eye, we can see all the way down to about a magnitude of 6. The brighter stars in the sky are around magnitude zero, with the brightest, Sirius, having a magnitude of about -1.5.

magnitude system | See magnitude.

main sequence | Abbreviated "MS". The region of a Hertzsprung-Russell diagram, running diagonally from hot and bright to cool and dim, stars that appear in this region derive their energy solely from hydrogen fusion. The MS contains roughly 90% of all stars.

main sequence fitting | Determining cosmic distances by comparing the main sequence regions in H-R diagrams of different star clusters.

main sequence star | A star that is actively fusing hydrogen into helium in its core. The inward gravitational force due to the mass of the star is balanced by the outward thermal pressure generated by nuclear fusion.

main sequence turn-off point | Refers to the point at which stars leave the main sequence in the H-R diagram as they exhaust the hydrogen in their core.

mass number | The total number of protons and neutrons in an atom.

masses | A measure of the inertia of an object and also of the strength of its gravitational interactions, with larger masses having greater inertia and stronger gravitational interactions. Mass is related to how much "stuff"—in the form of protons and neutrons—an object is made of, and is only changed by changing the amount of this stuff.

meteor | The flash of light we see when a solid object falls into our atmosphere and disintegrates. These objects vary in size from as small as sand to many meters in diameter. The largest reach the ground before burning up and create impact craters when they land. The pieces of rock or metal that remain are called meteorites.

meteor shower | When we see a large number of meteors in a relatively short time, created when Earth passes through a cloud of dust left over from the pass of a comet. For example, during its peak, observers might notice one or two meteors a minute from the Perseid meteor shower that occurs each year around August 12.

meter | Abbreviated "m". The fundamental SI unit of length, defined as the length of the path traveled by light in vacuum during a period of 1/299,792,458 s. A unit of length equal to about 39 inches. A kilometer is equal to 1000 meters.

metric system | See International System of Units. The metric system uses Celsius rather than Kelvin for temperature, but is otherwise the same as the International System of Units.

MeV | Megaelectron volt, or 1 million electron volts.

micro (prefix) | One-millionth; 10^{-6}. Denoted as μ (Greek lowercase mu), e.g., μm.

microwave | A region of the electromagnetic spectrum between infrared and radio. The energy of microwaves is a bit higher than radio waves. Their wavelengths are therefore shorter and are typically measured in centimeters (cm or 10^{-2} m).

Milky Way | Common name for the galaxy in which our Solar System is located.

milli (prefix) | One-thousandth; 10^{-3}. Denoted as m, e.g., mm.

model | A simplified explanation of how a natural system works that is based on empirical evidence and logic. To be useful, a model should make testable predictions. See also theory.

molecular cloud | A giant region of diffuse gases that can be several hundred light years across. They are composed mostly of molecular hydrogen, with helium and a few other elements dispersed throughout. Internal gravitation in colder denser regions of the cloud can trigger collapse and star formation. In addition to molecular hydrogen, molecular clouds contain molecules like CO, CH_4, NH_3, HCN, CH_2O and others.

molecule | Two or more atoms held together by chemical bonding.

moon | A celestial body that orbits a planet or smaller body. See also satellite.

moving cluster method | A method for determining the distances to clusters of stars that employs geometry and trigonometry to determine the cluster distances. This method would be useful just outside the boundary of using the parallax method.

multi-wavelength astronomy | The study of the Universe in all ranges of the electromagnetic spectrum, from radio waves to gammarays.

muon | a subatomic particle similar to, but more massive than, an electron.

natural satellite | See moon.

nebula | Plural, nebulae. Interstellar clouds of dust and gas, from the Greek, for *cloud.*

nebulae | Singular, nebula. Interstellar clouds of dust and gas, from the Greek, for *cloud*.

neutrino | An elementary particle that has an extremely small mass and only very weakly interacts with matter. The neutrino is part of the lepton family of particles. The majority of neutrinos detected on Earth come from the Sun.

neutron | One of the particles that makes up the nucleus (center) of atoms and has no charge. Neutrons are composed of three quarks.

neutron star | The collapsed core of a massive star, composed mostly of neutrons. Neutron stars are very small, with a diameter of about 10 kilometers. They have an enormous mass for their size, ranging from 1.4 solar masses to a bit more than twice that.

Newton's constant | See Universal Constant of Gravitation.

nuclear fission | The process by which heavy elements split apart into lighter ones. For instance, uranium nuclei can be split into two nuclei, each of which is roughly half the mass of the original uranium nucleus.

nuclear fusion | The process by which lighter elements like hydrogen and helium fuse together to make heavier elements like lithium, carbon, oxygen, etc.

nuclear reaction | The process by which the nucleus of an atom gains or loses neutrons and protons.

nuclei | Singular, *nucleus*. The central core of an atom, composed of neutrons and protons.

nucleus | Plural, nuclei. The central core of an atom, composed of neutrons and protons.

observatory | A facility that includes a telescope, either on the ground or in space.

Oort cloud | A region of space where long period comets originate, approximately 50,000 AU from the Sun.

open cluster | Loosely associated stars, numbering in the hundreds, that have formed together in the same cloud, but have not yet had time to drift apart.

open universe | A universe with no dark energy that does not contain enough mass to counteract its expansion; Omega is less than 1.

optical | The band of electromagnetic radiation that we can see with our eyes. It is intermediate in terms of energy and wavelength, between ultraviolet and infrared. Wavelengths range from approximately 400 to 750 nm, and energies are about one eV.

orbit | The path followed by a moon, planet, artificial satellite or other body, as dictated by gravity.

osmological constant | A constant term that can be added to Einstein's equations; works in the opposite direction to the gravity due to mass-energy; causes space to expand rather than contract.

oxidation | The combination of a chemical element with oxygen.

parallax | The apparent shift in position of a relatively nearby object compared to a more distant background as the location of the observer changes. Astronomically, it is half the angle that a star appears to move as Earth orbits from one side of the Sun to the other.

parsec | A unit of distance used by astronomers. An object one parsec away will exhibit a parallax of one arcsecond. One parsec equals about 3.3 light-years.

particle | See subatomic particle.

period | Time required for cyclic motion to repeat. For instance, the period of Earth to turn once around its axis is 24 hours, while the period for Earth to travel once around the Sun is 365 days. We would say that Earth has a *rotation period* of 24 hours and an *orbital period* of 365 days.

photo-excitation | The process by which an electron in an atom absorbs the energy from a photon and is excited to a higher energy state.

photoelectric effect | An effect whereby materials are induced to emit electrons when light shines onto them. The effect was explained in 1905 by Einstein by employing a particle theory of light.

photometry | The measurement of the brightness of astronomical objects. A standard result of photometry is the light curve (a plot of brightness versus time).

photon | A quantum (particle) of light or electromagnetic energy. Photons are have zero rest-mass and no electric charge.

pico (prefix) | One-trillionth; 10^{-12}. Denoted as p, e.g., pm.

Planck spectrum | See blackbody spectrum.

Planck's constant | A fundamental physical constant denoted by h. It has the value 6.626196×10^{-34} J s.

planet | Meaning "wanderer" in Greek, a celestial body that is massive enough for its own gravity to form itself as a spheroid but is not massive enough to begin thermonuclear fusion. Planets orbit a star or stellar remnant and have cleared their orbital paths of debris.

planetary nebula | Plural, nebulae. The expelled outer layers of low-mass stars, ionized by the ultraviolet radiation of a central white dwarf.

planetary nebulae | Singular, nebula. The expelled outer layers of low-mass stars, ionized by the ultraviolet radiation of a central white dwarf.

plasma | A gas that contains charged particles. It is composed of electrons that have been stripped from atoms, and the resulting positively charged particles called ions.

precision | The expected range of uncertainty of a physical measurement. Repeatability of that measurement. Precision differs from accuracy.

Principle of Invariance | this principle states that the spacetime separation of two events is a constant in Special Relativity, or in other words, that it is the same for all inertial frames of reference.

proper motion | The angular change in position of an astronomical object over time as seen from Earth. Measured in arcseconds per year.

proton | One of three subatomic particles that make up an atom. Protons are positively charged and located in the nucleus of an atom. Protons are composed of three quarks.

protostar | A young star that is still accreting matter from an accretion disk and that is enshrouded in a cloud of gas. A protostar is the earliest stage of a star's life, before it has even grown large enough to start nuclear fusion in its core.

pulsar | A type of magnetized spinning neutron star that emits a flash of light from a bright spot, like a lighthouse. Since the bright spot on the star's surface only points at us some of the time, it looks to us like it is pulsing on and off, hence the name pulsar.

Pythagorean Theorem | a theorem that relates the sides of a right triangle, stating that the square of the hypotenuse of the triangle (the side opposite the right angle) is equal to the sum of the squares of the other two sides. The theorem takes its name from the Ancient Greek philosopher and mathematician Pythagoras.

quantized | Discrete. For example, electrons in atoms, rather than having continuous energies, can only have a set of discrete or quantized energies, but not others.

quantum | Plural, quanta. A discrete minimal unit that is valid for physical systems. Photons, for example, are quanta of light.

quantum mechanics | The branch of physics that deals with the properties and behaviors of atoms and subatomic particles.

quantum system | A system that must be analyzed using quantum mechanics.

quark | A type of elementary particle that combines to make neutrons, protonsand other types of particles. There are six types of quarks, along with their anti-particle pairs. Three quarks combine to make neutrons and protons.

quintessence | A hypothetical form of non-constant dark energy postulated as an explanation of the observation of an accelerating rate of expansion of the Universe; involves a decaying energy field.

radial velocity | The velocity of an object along the observer's line of sight.

radian | A unit of angular measure, 1 radian = 57.3 degrees.

radiation | Energy emitted in the form of waves (example: light) or particles (example: electrons).

radio waves | The name given to the lowest energy region of the electromagnetic spectrum. Radio waves have wavelengths of meters (m), or even kilometers (km).

radioactive | An atom that is unstable and will break apart to become a new element, releasing energetic particles.

radioactivity | The natural or artificial process by which the nucleus of an atom is unstable and thereby breaks apart (decays) to become a new element. The decay process is accompanied by emission of energetic particles.

redshift | This is the name given to the apparent change in the wavelength of light due to the Doppler effect. Scientists know what the regular spectrum of a galaxy should look like (based on the spectrum of light emitted from known elements). If the light waves from a galaxy appear to have shifted towards higher frequency (blue), it is moving towards us, and if they have shifted toward a lower frequency (red), it means the object is moving away.

redshift (cosmological) | This is the name given to the apparent change in the wavelength of light due to the expansion of the Universe. The cosmological redshift is denoted by the letter z, and it is defined such that the Universe has expanded by an amount 1+z over the time the light has traveled to us. So an object with redshift z=1 is seen when the Universe was half its present size (it is twice as big as when the light was emitted), if z=2 the Universe is three times bigger than when the light was emitted, if z=3 the Universe is four times bigger, and so on.

reference frame | see frame.

relativistic | systems in which relativity is important, generally because the velocities are an appreciable fraction of the speed of light.

relativistic gamma | The factor in special relativity by which time and space are stretched or compressed by relative motion. Also, the ratio of the energy of a particle to its rest energy. The gamma factor depends on the relative velocity between reference frames approximately equal to x for small angles x (when x is measured in radians).

resolution | The fine-ness of a measurement. For example, a camera with a high resolution has the capability of capturing a more detailed image than a camera with a lower resolution.

rest energy | See rest mass.

rest energy | the energy a particle has in its rest frame, which depends on the particle's mass and the speed of light.

rest frame | a frame that is not moving with respect to an observer making measurements in it. We say that the observer is at rest in such a frame.

rest mass | The mass of an object measured when it is at rest relative to the observer measuring it.

rotate | turns on its own axis

satellite | A natural or man-made object that orbits a planet or other object.

scale | A ratio between the measurement of an object or event and its representation within a model. The scale can be represented as 1:X, where one is "one unit on the model" and is equal to X units in the actual system. On some maps, there is a scale that reads "one inch equals 10 miles" or something similar.

scientific notation | Using base-ten exponential form, e.g. 2.04×10^4 kg for 20,400 kg, to write numbers, especially very large or very small numbers.

second | Abbreviated "s". The fundamental SI unit of time, defined as the period of time equal to the duration of 9,192,631,770 periods of the radiation corresponding to the transition between two hyperfine levels of the ground state of the cesium-133 atom. There are 60 seconds in each minute and 3600 seconds in each hour of time.

Second Law of Thermodynamics | a physical law that says that the total entropy of a system must always increase.

SI Units | Abbreviated SI Units, also called the metric system, it is the scientific standard of carefully defined units of measurement. The SI unit of length is the meter, of time is the second, and of mass is the kilogram. Many other units are used to measure other quantities.

singularity | A place in the center of a black hole where the equations describing the mass density and gravitational force become infinite.

sky brightness | Extraneous light in the sky that is a result of scattered light from ground sources, light emitted from the atmosphere itself (perhaps due to its reaction with cosmic rays), and light from unresolved background sources.

small angle approximation | A mathematical approximation which amounts to tan(x) and sin(x) being approximately equal to x for small angles x (when x is measured in radians).

small angle formula | See small angle approximation.

SNR | See supernova remnant.

solar wind | The term given to the stream of charged particles that are ejected from the Sun's atmosphere. One of the effects of the interaction of solar wind and Earth's atmosphere is a creation of beautiful light patterns in sky known as auroras.

solar/star system | Defined as a system of celestial objects such as planets and asteroids orbiting one or more stars. When capitalized (Solar System), indicates the system associated with Earth and the Sun.

sources of error | Factors that can introduce errors into measurements and experiments. One of the most common sources of error originates in the instruments themselves.

spacetime | the four-dimensional system employed in special relativity that merges three spatial dimensions with one dimension of time and describes all events as points in that system: *(t, x, y, z)*. Three coordinates, *x, y,* and zdescribe the position of an event in space, and one, *t*, describes its position in time.

spacetime diagram | a simplified schematic representation of spacetime that generally shows the time axis as vertical with one axis of space being horizontal. The other two spatial dimensions are usually suppressed for simplicity. The diagrams are useful for understanding the relationship between events as seen by different observers in special relativity.

Special Theory of Relativity | Commonly called special relativity, the theory that predicts the behavior of objects moving at speeds close to the speed of light. One assumption of Special Relativity is that the speed of light in a vacuum is always constant.

spectra | Singular, *spectrum*. The distribution of intensity (i.e., number) of photons as a function of energy. Equivalently, it could be photon intensity distribution versus either wavelength or frequency.

Spectral Class | See stellar spectral classification.

spectrograph | Scientific instrument used to measure a spectrum.

spectroscopy | The scientific technique in which the intensity of light of different colors or wavelengths is measured. Comparing the measurements at different wavelengths can help to determine, for example, which elements are present in the light source.

spectrum | Plural, spectra. The distribution of intensity (i.e., number) of photons as a function of energy. Equivalently, it could be photon intensity distribution versus either wavelength or frequency.

speed | The distance (d) covered by a moving object in a given time (t). Mathematically, speed, s, is given by s = d/t. Differs from velocity in that velocity takes account of direction, not just how fast something moves.

speed of light | In a vacuum, denoted as "c", the speed of light is 299,792,458 m/s in all frames of reference, regardless of their relative states of motion.

spin | The quantum mechanical property of a particle that is analogous to the classical angular momentum (like a spinning top). For particles, the spin is always quantized and is either integral (for bosons) or half-integral (for fermions).

spiral arms | See spiral galaxy.

spiral galaxy | A galaxy whose primary feature is a flattened disk with long spiraling arms that extend out from a central core or bulge. The arms contain many massive young stars, a sign of high rates of star formation there.

standard candle | A celestial object that has a known inherent luminosity. Because these objects have a known luminosity output, their distances can be measured using their apparent brightness and the fact that brightness of an object falls as the inverse-square of distance from the object.

standard ruler | An astronomical object whose physical size is known. Using the known size and the small angle approximation, it is possible to determine the distance to the object.

stellar nurseries | Giant molecular clouds in galaxy, in which the density and temperature of the gas are such that parts of the clouds collapse under their own gravity to form new stars.

stellar spectral classification | A system in which stars are given a classification of O, B, A, F, G, A, K, or M based on their pattern of spectral absorption lines, which is related to their surface temperatures. O stars are the hottest and M stars are the coolest.

strong nuclear force | The force between quarks that keeps protons and neutrons bound within the nucleus. The force is also responsible for binding the nuclei themselves.

subatomic particles | Particles smaller than an atom, such as neutrons, protons and electrons, as well as other smaller particles like quarks, neutrinos, etc.

supermassive black hole | A black hole with mass on the order of millions or billions of solar masses. There is strong evidence that all large galaxies contain such black holes in their cores, including our own Milky Way, which contains a 4-million solar mass black hole at its center.

supernova | Plural, supernovae. The explosive collapse of the core of an evolved star. In massive stars, this collapse forms a neutron star or black hole. Under some circumstances low mass stars can also undergo supernovae as the result of the explosion of a particular type of white dwarf.

supernova remnant | The gas expelled during a supernova explosion, as well as the material swept up by that gas.

supernovae | Singular, *supernova.* The explosive collapse of the core of an evolved star. In massive stars, this collapse forms a neutron star or black hole. Under some circumstances low mass stars can also undergo supernovae as the result of the explosion of a particular type of white dwarf.

tangential velocity | In astronomy, the velocity of a star perpendicular to our line of sight, i.e., its velocity in the plane of the sky.

telescope | Optical instruments used to see great distances. Although originally telescopes were handheld, today they include both ground- and space-based varieties. Some examples are 10-meter motor-driven optical instruments, such as the Keck telescopes in Hawaii; the 27 antenna Very Large Array (VLA) radio observatory in New Mexico; and the orbiting Fermi Gamma ray Space Telescope some 550 km above Earth.

temperature | Most commonly a measure of the average energy of a particle in a system. Measured in kelvin (K) in the SI system.

Tera (prefix) | A trillion; 10^{12}. Denoted as T, e.g., TeV.

terrestrial | "Earth-like" planets. They are made mostly of rock, have solid surfaces, and typically have a mass comparable to Earth's. They generally have only one or two moons, if any. From the Latin, *terra*, meaning Earth.

theory | A conceptual framework that has explanatory and predictive power related to some aspect of the world. Theories generally encapsulate many experimental results and observations of the world into a coherent logical structure that makes testable predictions about related phenomena. For instance, Newton's theory of gravity explains the motions of the moon and falling objects on Earth, and makes predictions about the motions of other planets in the solar system as well as stars and galaxies.

time | A measure of how long it takes something to happen (an event's duration).

time dilation | The slowing of clocks that are in motion relative to an observer when compared to clocks at rest with respect to the observer. Predicted by Einstein's Special Theory of Relativity.

time dilation | the distortion of time between inertial frames. Observers in different inertial frames will measure time to pass more quickly in their own frame than in any inertial frame moving relative to their own.

Trans-Neptunian Objects | Trans-Neptunian Objects—Minor planets that orbit the Sun interior to the Kuiper Belt.

transit | When a planet passes in between Earth and the Sun such that the planet is seen to cross the face of the Sun. Only Mercury and Venus can undergo transits. Also, the passage of an extrasolar planet across the face of its star.

turbulent flow | Chaotic or disorganized motions within a fluid.

Type 1A supernova | The explosion of a carbon-oxygen (C-O) white dwarf that has accumulated enough material from a companion star to exceed the Chandrasekhar mass limit.

ultraviolet (UV) light | Electromagnetic radiation intermediate between the blue/violet end of visible light and x-rays. Ultraviolet radiation is more energetic than visible light but less than x-rays. Ultraviolet wavelengths are typically about 100 to 3800 nanometers (nm, 10^{-9} m).

uncertainty | The range of likely errors based on measurements of a given quantity, generally denoted as plus/minus (±) error-range and depending upon the experiment and measurement techniques. For example, a measurement listed as 6 ± 1 mm might be expected to have a true value of anywhere from 5 to 7 mm.

uncertainty principle | The position and the velocity of an object cannot both be known to perfect precision simultaneously; similarly, the energy and lifetime of virtual particles cannot both be known to perfect precision simultaneously.

uniform motion | Moving with a constant speed and direction.

Universal Constant of Gravitation | Denoted as capital G. The constant of proportionality in Newton's law of universal gravitation. It plays an analogous role in Einstein's general relativity. It is equal to 6.67428×10^{-11} m^3 / kg-sec^2.

Universe | Everything that exists, including Earth, planets, stars, galaxies, and all that they contain; the entire cosmos.

vacuum energy | The energy content of "empty" space; one possible explanation for the cosmological constant.

variable | A value that can change. For instance, the brightness of a pulsar changes depending on whether or not its beam of light is pointing toward us when we are looking at it, so its brightness is variable. Often in mathematical expressions, some parameters are allowed to change, and are thus variable.

velocity | How fast an object moves in a given direction, i.e., the speed of an object in a given direction. Velocity differs from speed because speed is how fast something moves without regard to direction.

visible light | Electromagnetic radiation at wavelengths which the human eye can see. We perceive this radiation as colors ranging from red (longer wavelengths; ~ 700 nanometers) to violet (shorter wavelengths; ~400 nanometers.) Also called optical light.

wavelength | The distance between adjacent peaks in a series of periodic waves. Also see electromagnetic spectrum.

worldline | The path taken by a particle through spacetime. This line connects all the events in the particle's history and future.

x-ray | High-energy electromagnetic radiation. X-rays are more energetic than ultraviolet light but less energetic than gamma rays. The energy of x-rays ranges roughly from 1 keV up to a few hundred keV. Their wavelengths are from about 10 nm down to about 10 pm.

± | "Plus/Minus," indicates the range of uncertainty of a value, e.g. 10.2 ± 0.4 kg indicates the mean value of the experiment was 10.2 kg, but there is a possibility that the true mass lies somewhere between 9.8—10.6 kg.

Detailed Licensing

Overview

Title: Physics 156: A Cosmology Workbook

Webpages: 71

All licenses found:

- CC BY 4.0: 94.4% (67 pages)
- Undeclared: 5.6% (4 pages)

By Page

- Physics 156: A Cosmology Workbook - *CC BY 4.0*
  - Front Matter - *CC BY 4.0*
    - TitlePage - *CC BY 4.0*
    - InfoPage - *CC BY 4.0*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
    - About the Authors - *CC BY 4.0*
  - 1: Workbook - *CC BY 4.0*
    - 1.1: Overview - *CC BY 4.0*
    - 1.2: Spacetime Geometry - *CC BY 4.0*
    - 1.3: Redshifts - *CC BY 4.0*
    - 1.4: Spatially Homogeneous and Isotropic Spacetimes - *CC BY 4.0*
    - 1.5: Euclidean Geometry - *CC BY 4.0*
    - 1.6: Distances as Determined by Standard Candles - *CC BY 4.0*
    - 1.7: The Distance-Redshift Relation - *CC BY 4.0*
    - 1.8: Dynamics of the Expansion - *CC BY 4.0*
    - 1.9: A Newtonian Homogeneous Expanding Universe - *CC BY 4.0*
    - 1.10: The Friedmann Equation - *CC BY 4.0*
    - 1.11: Particle Kinematics in an Expanding Universe - Newtonian Analysis - *CC BY 4.0*
    - 1.12: The Evolution of Mass-Energy Density and a First Glance at the Contents of the Cosmos - *CC BY 4.0*
    - 1.13: Energy and Momentum Conservation - *CC BY 4.0*
    - 1.14: Pressure and Energy Density Evolution - *CC BY 4.0*
    - 1.15: Distance and Magnitude - *CC BY 4.0*
    - 1.16: Parallax, Cepheid Variables, Supernovae, and Distance Measurement - *CC BY 4.0*
    - 1.17: Cosmological Data Analysis - *CC BY 4.0*
    - 1.18: The Early Universe - *CC BY 4.0*
    - 1.19: Equilibrium Statistical Mechanics - *CC BY 4.0*
      - 1.19.1: Chapter 19 footnotes - *CC BY 4.0*
    - 1.20: Equilibrium Particle Abundances - *CC BY 4.0*
    - 1.21: Hot and Cold Relics of the Big Bang - *CC BY 4.0*
    - 1.22: Overview of Thermal History - *CC BY 4.0*
    - 1.23: Big Bang Nucleosynthesis - Predictions - *CC BY 4.0*
    - 1.24: Big Bang Nucleosynthesis - Observations - *CC BY 4.0*
    - 1.25: Introduction to the Cosmic Microwave Background - *CC BY 4.0*
    - 1.26: The Spectrum of the CMB - *CC BY 4.0*
    - 1.27: Cosmic Microwave Background Anisotropies - *CC BY 4.0*
    - 1.28: Solving the Wave Equation with Fourier Transforms - *CC BY 4.0*
    - 1.29: The First Few Hundred Thousand Years- The Dynamics of the Primordial Plasma - *CC BY 4.0*
    - 1.30: Structure Formation - *CC BY 4.0*
    - 1.31: Galaxy Formation - *CC BY 4.0*
    - 1.32: Euclidean Geometry - *CC BY 4.0*
    - 1.33: Curvature - *CC BY 4.0*
    - 1.34: Galilean Relativity - *CC BY 4.0*
    - 1.35: Einstein Relativity - *CC BY 4.0*
    - 1.36: The Simplest Expanding Spacetime - *CC BY 4.0*
    - 1.37: Redshifts - *CC BY 4.0*
  - 2: Exercise Solutions - *CC BY 4.0*
    - 2.1: S01. Euclidean Geometry - SOLUTIONS - *CC BY 4.0*
    - 2.2: S02. Curvature - SOLUTIONS - *CC BY 4.0*
    - 2.3: S03. Galilean Relativity - SOLUTIONS - *CC BY 4.0*
    - 2.4: S04. Einstein Relativity - SOLUTIONS - *CC BY 4.0*
    - 2.5: S05. Spacetime - SOLUTIONS - *CC BY 4.0*
    - 2.6: S06. Redshifts - SOLUTIONS - *CC BY 4.0*
    - 2.7: S07. Distances as Determined by Standard Candles - SOLUTIONS - *CC BY 4.0*
    - 2.8: S08. The Distance-Redshift Relation - SOLUTIONS - *CC BY 4.0*

- 2.9: S10. A Newtonian Homogeneous Expanding Universe - SOLUTIONS - *CC BY 4.0*
- 2.10: S11. The Friedmann Equation - SOLUTIONS - *CC BY 4.0*
- 2.11: S12. Particle Kinematics in an Expanding Universe- Newtonian Analysis - SOLUTIONS - *CC BY 4.0*
- 2.12: S13. The Evolution of Mass-Energy Density and a First Glance at the Contents of the Cosmos - SOLUTIONS - *CC BY 4.0*
- 2.13: S15 Pressure and Energy Density Evolution SOLUTIONS - *CC BY 4.0*
- 2.14: S16. Distance and Magnitude - SOLUTIONS - *CC BY 4.0*
- 2.15: S17. Parallax, Cepheid Variables, Supernovae, and Distance Measurement - SOLUTIONS - *CC BY 4.0*
- 2.16: S20 Equilibrium Statistical Mechanics SOLUTIONS - *CC BY 4.0*
- 2.17: Equilibrium Particle Abundances - *CC BY 4.0*
- 2.18: S22. Hot and Cold Relics of the Big Bang - *CC BY 4.0*
- 2.19: S24. Big Bang Nucleosynthesis- Predictions - *CC BY 4.0*

- Back Matter - *CC BY 4.0*
  - Index - *CC BY 4.0*
  - Glossary - *CC BY 4.0*
  - Detailed Licensing - *Undeclared*
  - Detailed Licensing - *Undeclared*

Detailed Licensing

Overview

Title: Physics 156: A Cosmology Workbook

Webpages: 71

All licenses found:

- CC BY 4.0: 94.4% (67 pages)
- Undeclared: 5.6% (4 pages)

By Page

- Physics 156: A Cosmology Workbook - *CC BY 4.0*
  - Front Matter - *CC BY 4.0*
    - TitlePage - *CC BY 4.0*
    - InfoPage - *CC BY 4.0*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
    - About the Authors - *CC BY 4.0*
  - 1: Workbook - *CC BY 4.0*
    - 1.1: Overview - *CC BY 4.0*
    - 1.2: Spacetime Geometry - *CC BY 4.0*
    - 1.3: Redshifts - *CC BY 4.0*
    - 1.4: Spatially Homogeneous and Isotropic Spacetimes - *CC BY 4.0*
    - 1.5: Euclidean Geometry - *CC BY 4.0*
    - 1.6: Distances as Determined by Standard Candles - *CC BY 4.0*
    - 1.7: The Distance-Redshift Relation - *CC BY 4.0*
    - 1.8: Dynamics of the Expansion - *CC BY 4.0*
    - 1.9: A Newtonian Homogeneous Expanding Universe - *CC BY 4.0*
    - 1.10: The Friedmann Equation - *CC BY 4.0*
    - 1.11: Particle Kinematics in an Expanding Universe - Newtonian Analysis - *CC BY 4.0*
    - 1.12: The Evolution of Mass-Energy Density and a First Glance at the Contents of the Cosmos - *CC BY 4.0*
    - 1.13: Energy and Momentum Conservation - *CC BY 4.0*
    - 1.14: Pressure and Energy Density Evolution - *CC BY 4.0*
    - 1.15: Distance and Magnitude - *CC BY 4.0*
    - 1.16: Parallax, Cepheid Variables, Supernovae, and Distance Measurement - *CC BY 4.0*
    - 1.17: Cosmological Data Analysis - *CC BY 4.0*
    - 1.18: The Early Universe - *CC BY 4.0*
    - 1.19: Equilibrium Statistical Mechanics - *CC BY 4.0*
      - 1.19.1: Chapter 19 footnotes - *CC BY 4.0*
    - 1.20: Equilibrium Particle Abundances - *CC BY 4.0*
    - 1.21: Hot and Cold Relics of the Big Bang - *CC BY 4.0*
    - 1.22: Overview of Thermal History - *CC BY 4.0*
    - 1.23: Big Bang Nucleosynthesis - Predictions - *CC BY 4.0*
    - 1.24: Big Bang Nucleosynthesis - Observations - *CC BY 4.0*
    - 1.25: Introduction to the Cosmic Microwave Background - *CC BY 4.0*
    - 1.26: The Spectrum of the CMB - *CC BY 4.0*
    - 1.27: Cosmic Microwave Background Anisotropies - *CC BY 4.0*
    - 1.28: Solving the Wave Equation with Fourier Transforms - *CC BY 4.0*
    - 1.29: The First Few Hundred Thousand Years- The Dynamics of the Primordial Plasma - *CC BY 4.0*
    - 1.30: Structure Formation - *CC BY 4.0*
    - 1.31: Galaxy Formation - *CC BY 4.0*
    - 1.32: Euclidean Geometry - *CC BY 4.0*
    - 1.33: Curvature - *CC BY 4.0*
    - 1.34: Galilean Relativity - *CC BY 4.0*
    - 1.35: Einstein Relativity - *CC BY 4.0*
    - 1.36: The Simplest Expanding Spacetime - *CC BY 4.0*
    - 1.37: Redshifts - *CC BY 4.0*
  - 2: Exercise Solutions - *CC BY 4.0*
    - 2.1: S01. Euclidean Geometry - SOLUTIONS - *CC BY 4.0*
    - 2.2: S02. Curvature - SOLUTIONS - *CC BY 4.0*
    - 2.3: S03. Galilean Relativity - SOLUTIONS - *CC BY 4.0*
    - 2.4: S04. Einstein Relativity - SOLUTIONS - *CC BY 4.0*
    - 2.5: S05. Spacetime - SOLUTIONS - *CC BY 4.0*
    - 2.6: S06. Redshifts - SOLUTIONS - *CC BY 4.0*
    - 2.7: S07. Distances as Determined by Standard Candles - SOLUTIONS - *CC BY 4.0*
    - 2.8: S08. The Distance-Redshift Relation - SOLUTIONS - *CC BY 4.0*

- 2.9: S10. A Newtonian Homogeneous Expanding Universe - SOLUTIONS - *CC BY 4.0*
- 2.10: S11. The Friedmann Equation - SOLUTIONS - *CC BY 4.0*
- 2.11: S12. Particle Kinematics in an Expanding Universe- Newtonian Analysis - SOLUTIONS - *CC BY 4.0*
- 2.12: S13. The Evolution of Mass-Energy Density and a First Glance at the Contents of the Cosmos - SOLUTIONS - *CC BY 4.0*
- 2.13: S15 Pressure and Energy Density Evolution SOLUTIONS - *CC BY 4.0*
- 2.14: S16. Distance and Magnitude - SOLUTIONS - *CC BY 4.0*
- 2.15: S17. Parallax, Cepheid Variables, Supernovae, and Distance Measurement - SOLUTIONS - *CC BY 4.0*
- 2.16: S20 Equilibrium Statistical Mechanics SOLUTIONS - *CC BY 4.0*
- 2.17: Equilibrium Particle Abundances - *CC BY 4.0*
- 2.18: S22. Hot and Cold Relics of the Big Bang - *CC BY 4.0*
- 2.19: S24. Big Bang Nucleosynthesis- Predictions - *CC BY 4.0*

- Back Matter - *CC BY 4.0*
  - Index - *CC BY 4.0*
  - Glossary - *CC BY 4.0*
  - Detailed Licensing - *Undeclared*
  - Detailed Licensing - *Undeclared*