

8.4: Cosmological Solutions (Part 2)

A Singularity at the Big Bang

The Friedmann equations only allow a constant a in the case where Λ is perfectly tuned relative to the other parameters, and even this artificially fine-tuned equilibrium turns out to be unstable. These considerations make a static cosmology implausible on theoretical grounds, and they are also consistent with the observed Hubble expansion ([section 8.2](#)).

Since the universe is not static, what happens if we use general relativity to extrapolate farther and farther back in time?

If we extrapolate the Friedmann equations backward in time, we find that they always have a $a = 0$ at some point in the past, and this occurs regardless of the details of what we assume about the matter and radiation that fills the universe. To see this, note that, as discussed in [example 14](#), radiation is expected to dominate the early universe, for generic reasons that are not sensitive to the (substantial) observational uncertainties about the universe's present-day mixture of ingredients. Under radiation-dominated conditions, we can approximate $\Lambda = 0$ and $P = \frac{\rho}{3}$ ([example 14](#)) in the first Friedmann equation, finding

$$\frac{\ddot{a}}{a} = -\frac{8\pi}{3}\rho \quad (8.2.4)$$

where ρ is the density of mass-energy due to radiation. Since $\frac{\ddot{a}}{a}$ is always negative, the graph of $a(t)$ is always concave down, and since a is currently increasing, there must be some time in the past when $a = 0$. One can readily verify that this is not just a coordinate singularity; the Ricci scalar curvature R^a_a diverges, and the singularity occurs at a finite proper time in the past.

In [section 6.3](#), we saw that a black hole contains a singularity, but it appears that such singularities are always hidden behind event horizons, so that we can never observe them from the outside. The FRW singularity, however, is not hidden behind an event horizon. It lies in our past light-cone, and our own world-lines emerged from it. The universe, it seems, originated in a Big Bang, a concept that originated with the Belgian Roman Catholic priest Georges Lemaître.



Figure 8.4.5 - Georges Lemaître (1894- 1966) proposed in 1927 that our universe be modeled in general relativity as a spacetime in which space expanded over time. Lemaître's ideas were initially treated skeptically by Eddington and Einstein, who told him, "Your calculations are correct, but your physics is abominable." Later, as Hubble's observational evidence for cosmological expansion became widely accepted, both Einstein and Eddington became converts, helping to bring Lemaître's ideas to the attention of the community. In 1931, an emboldened Lemaître described the idea that the universe began from a "Primeval Atom" or "Cosmic Egg." The name that eventually stuck was "Big Bang," coined by Fred Hoyle as a derisive term.

exercise 8.4.1:

Self-check: Why is it not correct to think of the Big Bang as an explosion that occurred at a specific point in space?

Does the FRW singularity represent something real about our universe?

One thing to worry about is the accuracy of our physical modeling of the radiation-dominated universe. The presence of an initial singularity in the FRW solutions does not depend sensitively on assumptions like $P = \frac{\rho}{3}$, but it is still disquieting that no laboratory experiment has ever come close to attaining the conditions under which we could test whether a gas of photons produces gravitational fields as predicted by general relativity. We saw in [section 8.1](#) that static electric fields do produce gravitational fields as predicted, but this is not the same as an empirical confirmation that electromagnetic waves also act as gravitational sources in exactly the manner that general relativity claims. We do, however, have a consistency check in the form of the abundances of nuclei. Calculations of nuclear reactions in the early, radiation-dominated universe predict certain abundances of hydrogen, helium, and deuterium. In particular, the relative abundance of helium and deuterium is a sensitive test of the relationships among a , \dot{a} , and \ddot{a} predicted by the FRW equations, and they confirm these relationships to a precision of about $5 \pm 4\%$.¹⁵

Note

Steigman, Ann. Rev. Nucl. Part. Sci. 57 (2007) 463. These tests are stated in terms of the Hubble “constant” $H = \frac{\dot{a}}{a}$, which is actually varying over cosmological time-scales. The nuclear helium-deuterium ratio is sensitive to $\frac{\ddot{H}}{H}$.

An additional concern is whether the Big Bang singularity is just a product of the unrealistic assumption of perfect symmetry that went into the FRW cosmology. One of the Penrose-Hawking singularity theorems proves that it is not.¹⁶ This particular singularity theorem requires three conditions: (1) the strong energy condition holds; (2) there are no closed timelike curves; and (3) a trapped surface exists in the past timelike geodesics originating at some point. The requirement of a trapped surface can fail if the universe is inhomogeneous to $\gtrsim 10^{-4}$, but observations of the cosmic microwave background rule out any inhomogeneity this large (see [section 8.2](#)). The other possible failure of the assumptions is that if the cosmological constant is large enough, it violates the strong energy equation, and we can have a Big Bounce rather than a Big Bang (see [section 8.2](#)).

An Exceptional Case: the Milne Universe

There is still a third loophole in our conclusion that the Big Bang singularity must have existed. Consider the special case of the FRW analysis, found by Milne in 1932 (long before FRW), in which the universe is completely empty, with $\rho = 0$ and $\Lambda = 0$. This is of course not consistent with the fact that the universe contains stars and galaxies, but we might wonder whether it could tell us anything interesting as a simplified approximation to a very dilute universe. The result is that the scale factor a varies linearly with time ([problem 3](#)). If a is not constant, then there exists a time at which $a = 0$, but this doesn’t turn out to be a real singularity (which isn’t surprising, since there is no matter to create gravitational fields). Let this universe have a scattering of test particles whose masses are too small to invalidate the approximation of $\rho = 0$, and let the test particles be at rest in the (r, θ, ϕ) coordinates. The linear dependence of a on t means that these particles simply move inertially and without any gravitational interactions, spreading apart from one another at a constant rate like the raisins in a rising loaf of raisin bread. The Friedmann equations require $k = -1$, so the spatial geometry is one of constant negative curvature.

The Milne universe is in fact flat spacetime described in tricky coordinates. The connection can be made as follows. Let a spherically symmetric cloud of test particles be emitted by an explosion that occurs at some arbitrarily chosen event in flat spacetime. Make the cloud’s density be nonuniform in a certain specific way, so that every observer moving along with a test particle (called a comoving observer) sees the same local conditions in his own frame; due to Lorentz contraction by a factor γ , this requires that the density be proportional to γ as described by the observer O who remained at the origin. This scenario turns out to be identical to the Milne universe under the change of coordinates from spatially flat coordinates (T, R) to FRW coordinates (t, r) , where $t = \frac{T}{\gamma}$ is the proper time and $r = v\gamma$. (Cf. [problem 12](#))

The Milne universe may be useful as an inoculation against the common misconception that the Big Bang was an explosion of matter spreading out into a preexisting vacuum. Such a description seems obviously incompatible with homogeneity, since, for example, an observer at the edge of the cloud sees the cloud filling only half of the sky. But isn’t this a logical contradiction, since the Milne universe *does* have an explosion into vacuum, and yet it was derived as a special case of the FRW analysis, which explicitly assumed homogeneity? It is not a contradiction, because a comoving observer never actually sees an edge. In the limit as we approach the edge, the density of the cloud (as seen by the observer who stayed at the origin) approaches infinity, and the

Lorentz contraction also approaches infinity, so that O considers them to be like Hamlet saying, “I could be bounded in a nutshell, and count myself a king of infinite space.” This logic only works in the case of the Milne universe. The explosion-into-preexisting-vacuum interpretation fails in Big Bang cosmologies with $\rho \neq 0$.

Observability of Expansion

Brooklyn is Not Expanding!

The proper interpretation of the expansion of the universe, as described by the Friedmann equations, can be tricky. The example of the Milne universe encourages us to imagine that the expansion would be undetectable, since the Milne universe can be described as either expanding or not expanding, depending on the choice of coordinates. A more general consequence of coordinate-independence is that relativity does not pick out any preferred distance scale. That is, if all our meter-sticks expand, and the rest of the universe expands as well, we would have no way to detect the expansion. The flaw in this reasoning is that the Friedmann equations only describe the average behavior of spacetime. As dramatized in the classic Woody Allen movie “Annie Hall:” “Well, the universe is everything, and if it’s expanding, someday it will break apart and that would be the end of everything!” “What has the universe got to do with it? You’re here in Brooklyn! Brooklyn is not expanding!”

To organize our thoughts, let’s consider the following hypotheses:

1. The distance between one galaxy and another increases at the rate given by $a(t)$ (assuming the galaxies are sufficiently distant from one another that they are not gravitationally bound within the same galactic cluster, supercluster, etc.).
2. The wavelength of a photon increases according to $a(t)$ as it travels cosmological distances.
3. The size of the solar system increases at this rate as well (i.e., gravitationally bound systems get bigger, including the earth and the Milky Way).
4. The size of Brooklyn increases at this rate (i.e., electromagnetically bound systems get bigger).
5. The size of a helium nucleus increases at this rate (i.e., systems bound by the strong nuclear force get bigger).

We can imagine that:

- All the above hypotheses are true.
- All the above hypotheses are false, and in fact none of these sizes increases at all.
- Some are true and some false

If all five hypotheses were true, the expansion would be undetectable, because all available meter-sticks would be expanding together. Likewise if no sizes were increasing, there would be nothing to detect. These two possibilities are really the same cosmology, described in two different coordinate systems. But the Ricci and Einstein tensors were carefully constructed so as to be intrinsic. The fact that the expansion affects the Einstein tensor shows that it cannot be interpreted as a mere coordinate expansion. Specifically, suppose someone tells you that the FRW metric can be made into a flat metric by a change of coordinates. (I have come across this claim on internet forums.) The linear structure of the tensor transformation equations guarantees that a nonzero tensor can never be made into a zero tensor by a change of coordinates. Since the Einstein tensor is nonzero for an FRW metric, and zero for a flat metric, the claim is false.

exercise 8.4.2

Self-check: The reasoning above implicitly assumed a non-empty universe. Convince yourself that it fails in the special case of the Milne universe.

We can now see some of the limitations of a common metaphor used to explain cosmic expansion, in which the universe is visualized as the surface of an expanding balloon. The metaphor correctly gets across several ideas: that the Big Bang is not an explosion that occurred at a preexisting point in empty space; that hypothesis 1 above holds; and that the rate of recession of one galaxy relative to another is proportional to the distance between them. Nevertheless the metaphor may be misleading, because if we take a laundry marker and draw any structure on the balloon, that structure will expand at the same rate. But this implies that hypotheses 1-5 all hold, which cannot be true.

Since some of the five hypotheses must be true and some false, and we would like to sort out which are which. It should also be clear by now that these are not five independent hypotheses. For example, we can test empirically whether the ratio of Brooklyn’s size to the distances between galaxies changes like $a(t)$, remains constant, or changes with some other time dependence, but it is only the ratio that is actually observable.

Empirically, we find that hypotheses 1 and 2 are true (i.e., the photon's wavelength maintains a constant ratio with the intergalactic distance scale), while 3, 4, and 5 are false. For example, the orbits of the planets in our solar system have been measured extremely accurately by radar reflection and by signal propagation times to space probes, and no expanding trend is detected.

General-relativistic Predictions

Does general relativity correctly reproduce these observations? General relativity is mainly a theory of gravity, so it should be well within its domain to explain why the solar system does not expand detectably while intergalactic distances do. It is impractical to solve the Einstein field equations exactly so as to describe the internal structure of all the bodies that occupy the universe: galaxies, superclusters, etc. We can, however, handle simple cases, as in [example 20](#), where we display an exact solution for the case of a universe containing only two things: an isolated black hole, and an energy density described by a cosmological constant. We find that the characteristic scale of the black hole, i.e., the radius of its event horizon, does not increase with time. A fuller treatment of these issues is given [later](#), after some facts about realistic cosmologies have been established. The result is that although bound systems like the solar system are in some cases predicted to expand, the expansion is absurdly small, too small to measure, and much smaller than the rate of expansion of the universe in general as represented by the scale factor $a(t)$. This agrees with observation.

It is easy to show that atoms and nuclei do not *steadily* expand over time. because such an expansion would violate either the equivalence principle or the basic properties of quantum mechanics. One way of stating the equivalence principle is that the local geometry of spacetime is always approximately Lorentzian, so that the laws of physics do not depend on one's position or state of motion. Among these laws of physics are the principles of quantum mechanics, which imply that an atom or a nucleus has a well-defined ground state, with a certain size that depends only on fundamental constants such as Planck's constant and the masses of the particles involved. Atoms and nuclei do experience deformation due to gravitational strains ([examples 24-25](#)), but these deformations do not increase with time, and would only be detectable if cosmological expansion were to accelerate radically ([example 26](#)).

This is different from the case of a photon traveling across the universe. The argument given above fails, because the photon does not have a ground state. The photon *does* expand, and this is required by the correspondence principle. If the photon did not expand, then its wavelength would remain constant, and this would be inconsistent with the classical theory of electromagnetism, which predicts a Doppler shift due to the relative motion of the source and the observer. One can choose to describe cosmological redshifts either as Doppler shifts or as expansions of wavelength due to cosmological expansion.

A nice way of discussing atoms, nuclei, photons, and solar systems all on the same footing is to note that in geometrized units, the units of mass and length are the same. Therefore the existence of any fundamental massive particle sets a universal length scale, one that will be known to any intelligent species anywhere in the universe. Since photons are massless, they can't be used to set a universal scale in this way; a photon has a certain mass-energy, but that mass-energy can take on any value. Similarly, a solar system sets a length scale, but not a universal one; the radius of a planet's orbit can take on any value. A universe without massive fundamental particles would be a universe without length measurement. It would obey the laws of conformal geometry, in which angles and light-cones were the only measures. This is the reason that atoms and nuclei, which are made of massive fundamental particles, do not expand.

More than One Dimension Required

Another good way of understanding why a photon expands, while an atom does not, is to recall that a one-dimensional space can never have any intrinsic curvature. If the expansion of atoms were to be detectable, we would need to detect it by comparing against some other meter-stick. Let's suppose that a hydrogen atom expands more, while a more tightly bound uranium atom expands less, so that over time, we can detect a change in the ratio of the two atoms' sizes. The world-lines of the two atoms are one-dimensional curves in spacetime. They are housed in a laboratory, and although the laboratory does have some spatial extent, the equivalence principle guarantees that to a good approximation, this small spatial extent doesn't matter. This implies an intrinsic curvature in a one-dimensional space, which is mathematically impossible, so we have a proof by contradiction that atoms do not expand steadily.

Now why does this one-dimensionality argument fail for photons and galaxies? For a pair of galaxies, it fails because the galaxies are not sufficiently close together to allow them both to be covered by a single Lorentz frame, and therefore the set of world-lines comprising the observation cannot be approximated well as lying within a one-dimensional space. Similar reasoning applies for cosmological redshifts of photons received from distant galaxies. One could instead propose flying along in a spaceship next to an electromagnetic wave, and monitoring the change in its wavelength while it is in flight. All the world-lines involved in such an

experiment would indeed be confined to a one-dimensional space. The experiment is impossible, however, because the measuring apparatus cannot be accelerated to the speed of light. In reality, the speed of the light wave relative to the measuring apparatus will always equal c , so the two world-lines involved in the experiment will diverge, and will not be confined to a one-dimensional region of spacetime.

Example 16: A cosmic girdle

Since cosmic expansion has no significant effect on Brooklyn, nuclei, and solar systems, we might be tempted to infer that its effect on any solid body would also be negligible. To see that this is not true, imagine that we live in a closed universe, and the universe has a leather belt wrapping around it on a closed spacelike geodesic. All parts of the belt are initially at rest relative to the local galaxies, and the tension is initially zero everywhere. The belt must stretch and eventually break: for if not, then it could not remain everywhere at rest with respect to the local galaxies, and this would violate the symmetry of the initial conditions, since there would be no way to pick the direction in which a certain part of the belt should begin accelerating.

Example 17: Østvang's quasi-metric relativity

Over the years, a variety of theories of gravity have been proposed as alternatives to general relativity. Some of these, such as the Brans-Dicke theory, remain viable, i.e., they are consistent with all the available experimental data that have been used to test general relativity. One of the most important reasons for trying to construct such theories is that it can be impossible to interpret tests of general relativity's predictions unless one also possesses a theory that predicts something different. This issue, for example, has made it impossible to test Einstein's century-old prediction that gravitational effects propagate at c , since there is no viable theory available that predicts any other speed for them (see [section 9.1](#)).

Østvang (arxiv.org/abs/gr-qc/0112025v6) has proposed an alternative theory of gravity, called quasi-metric relativity, which, unlike general relativity, predicts a significant cosmological expansion of the solar system, and which is claimed to be able to explain the observation of small, unexplained accelerations of the Pioneer space probes that remain after all accelerations due to known effects have been subtracted (the "Pioneer anomaly"). We've seen above that there are a variety of arguments against such an expansion of the solar system, and that many of these arguments do not require detailed technical calculations but only knowledge of certain fundamental principles, such as the structure of differential geometry (no intrinsic curvature in one dimension), the equivalence principle, and the existence of ground states in quantum mechanics. We therefore expect that Østvang's theory, if it is logically self-consistent, will probably violate these assumptions, but that the violations must be relatively small if the theory is claimed to be consistent with existing observations. This is in fact the case. The theory violates the strictest form of the equivalence principle.

Over the years, a variety of explanations have been proposed for the Pioneer anomaly, including both glamorous ones (a modification of the $\frac{1}{r^2}$ law of gravitational forces) and others more pedestrian (effects due to outgassing of fuel, radiation pressure from sunlight, or infrared radiation originating from the spacecrafts radioisotope thermoelectric generator). Calculations by Iorio¹⁷ in 2006-2009 show that if the force law for gravity is modified in order to explain the Pioneer anomalies, and if gravity obeys the equivalence principle, then the results are inconsistent with the observed orbital motion of the satellites of Neptune. This makes gravitational explanations unlikely, but does not obviously rule out Østvang's theory, since the theory is not supposed to obey the equivalence principle. Østvang says¹⁸ that his theory predicts an expansion of $\sim 1\text{m/yr}$ in the orbit of Triton's moon Nereid, which is consistent with observation. In December 2010, the original discoverers of the effect made a statement in the popular press that they had a new analysis, which they were preparing to publish in a scientific paper, in which the size of the anomaly would be drastically revised downward, with a far greater proportion of the acceleration being accounted for by thermal effects. In my opinion this revision, combined with the putative effect's violation of the equivalence principle, make it clear that the anomaly is not gravitational.

¹⁷ <http://arxiv.org/abs/0912.2947v1>

¹⁸ private communication, Jan. 4, 2010

Does Space Expand?

Finally, the balloon metaphor encourages us to interpret cosmological expansion as a phenomenon in which space itself expands, or perhaps one in which new space is produced. Does space really expand? Without posing the question in terms of more rigorously defined, empirically observable quantities, we can't say yes or no. It is merely a matter of which definitions one chooses and which

conceptual framework one finds easier and more natural to work within. Bunn and Hogg have stated the minority view against expansion of space¹⁹, while the opposite opinion is given by Francis et al.²⁰

As an example of a self-consistent set of definitions that lead to the conclusion that space does expand, Francis et al. give the following. Define eight observers positioned at the corners of a cube, at cosmological distances from one another. Let each observer be at rest relative to the local matter and radiation that were used as ingredients in the FRW cosmology. (For example, we know that our own solar system is *not* at rest in this sense, because we observe that the cosmic microwave background radiation is slightly Doppler shifted in our frame of reference.) Then these eight observers will observe that, over time, the volume of the cube grows as expected according to the cube of the function $a(t)$ in the FRW model.

This establishes that expansion of space is a plausible interpretation. To see that it is not the only possible interpretation, consider the following example. A photon is observed after having traveled to earth from a distant galaxy G, and is found to be red-shifted. Alice, who likes expansion, will explain this by saying that while the photon was in flight, the space it occupied expanded, lengthening its wavelength. Betty, who dislikes expansion, wants to interpret it as a kinematic red shift, arising from the motion of galaxy G relative to the Milky Way Galaxy, M. If Alice and Betty's disagreement is to be decided as a matter of absolute truth, then we need some objective method for resolving an observed redshift into two terms, one kinematic and one gravitational. But we've seen in [section 7.4](#) that this is only possible for a stationary spacetime, and cosmological spacetimes are not stationary: regardless of an observer's state of motion, he sees a change over time in observables such as density of matter and curvature of spacetime. As an extreme example, suppose that Betty, in galaxy M, receives a photon without realizing that she lives in a closed universe, and the photon has made a circuit of the cosmos, having been emitted from her own galaxy in the distant past. If she insists on interpreting this as a kinematic red shift, she must conclude that her galaxy M is moving at some extremely high velocity relative to itself. This is in fact not an impossible interpretation, if we say that M's high velocity is relative to itself *in the past*. An observer who sets up a frame of reference with its origin fixed at galaxy G will happily confirm that M has been accelerating over the eons. What this demonstrates is that we can split up a cosmological red shift into kinematic and gravitational parts in any way we like, depending on our choice of coordinate system (see also [section 7.5](#)).

Example 18: A cosmic whip

The cosmic girdle of example 16 does not transmit any information from one part of the universe to another, for its state is the same everywhere by symmetry, and therefore an observer near one part of the belt gets no information that is any different from what would be available to an observer anywhere else.

Now suppose that the universe is open rather than closed, but we have a rope that, just like the belt, stretches out over cosmic distances along a spacelike geodesic. If the rope is initially at rest with respect to a particular galaxy G (or, more strictly speaking, with respect to the locally averaged cosmic medium), then by symmetry the rope will always remain at rest with respect to G, since there is no way for the laws of physics to pick a direction in which it should accelerate. Now the residents of G cut the rope, release half of it, and tie the other half securely to one of G's spiral arms using a square knot. If they do this smoothly, without varying the rope's tension, then no vibrations will propagate, and everything will be as it was before on that half of the rope. (We assume that G is so massive relative to the rope that the rope does not cause it to accelerate significantly.)

Can observers at distant points observe the tail of the rope whipping by at a certain speed, and thereby infer the velocity of G relative to them? This would produce all kinds of strange conclusions. For one thing, the Hubble law says that this velocity is directly proportional to the length of the rope, so by making the rope long enough we could make this velocity exceed the speed of light. We've also convinced ourselves that the relative velocity of cosmologically distant objects is not even well defined in general relativity, so it clearly can't make sense to interpret the rope-end's velocity in that way.

The way out of the paradox is to recognize that disturbances can only propagate along the rope at a certain speed v . Let's say that the information is transmitted in the form of longitudinal vibrations, in which case it propagates at the speed of sound. For a rope made out of any known material, this is far less than the speed of light, and we've also seen in [example 14](#) and in [problem 4](#) that relativity places fundamental limits on the properties of all possible materials, guaranteeing $v < c$. We can now see that all we've accomplished with the rope is to recapitulate using slower sound waves the discussion that was carried out earlier using light waves. The sound waves may perhaps preserve some information about the state of motion of galaxy G long ago, but all the same ambiguities apply to its interpretation as in the case of light waves — and in addition, we suspect that the rope has long since parted somewhere along its length.

References

¹⁶ Hawking and Ellis, “The Cosmic Black-Body Radiation and the Existence of Singularities in Our Universe,” *Astrophysical Journal*, 152 (1968) 25. Available online at articles.adsabs.harvard.edu.

¹⁹ <http://arxiv.org/abs/0808.1081v2>

²⁰ <http://arxiv.org/abs/0707.0380v1>

This page titled [8.4: Cosmological Solutions \(Part 2\)](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Benjamin Crowell](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.