# 7

# Cosmological implications and experimental bounds

In the previous chapters solitons were examined largely as theoretical constructs. Let us now address the question of whether they exist as actual physical objects. Condensed matter systems with structures analogous to kinks and vortices certainly exist and have been well studied; some of these have already been briefly mentioned. However, there has been as yet no confirmed experimental or observational evidence of a soliton in a relativistic quantum field theory. The natural question, then, is what conclusions can be drawn from this. The most plausible source of domain walls and strings is as relics surviving from the early universe. The same is true of magnetic monopoles if, as in grand unified theories, their masses are far beyond the reach of possible accelerator experiments. As we will see, all of these could have been produced during the course of symmetry-breaking cosmological phase transitions. Comparison of the expected production rates with the present-day bounds on the abundances of these objects yields important constraints on the underlying field theories and cosmological scenarios.

## 7.1 Brief overview of big bang cosmology

There is strong evidence, both from the spatial distribution of galaxies and, especially, observations of the cosmic microwave background radiation, that the universe (or at least the part accessible to our observations) possesses a high degree of spatial homogeneity and isotropy. Any homogeneous and isotropic spacetime can be described by the Robertson–Walker metric, which can be written as

$$ds^2 = dt^2 - a(t)^2 \left( \frac{dr^2}{1 - kr^2} + r^2 d\theta^2 + r^2 \sin\theta^2 d\varphi^2 \right). \tag{7.1}$$

Here $k$ indicates the nature of the spatial slices. It has three possible values, yielding flat Euclidean space ($k = 0$), a three-dimensional sphere ($k = 1$), or a three-dimensional hyperboloid ($k = -1$). These are referred to as flat, closed,

and open universes, respectively. In an open or a closed universe the scale factor $a(t)$ is the time-dependent curvature radius. For a flat universe the overall scale of $a$ is arbitrary, but the ratio of its values at two different times is a measure of the cosmic expansion and is physically meaningful.

The coordinates used here are comoving coordinates. A worldline with fixed $r$, $\theta$, and $\varphi$ is a geodesic, with $t$ measuring the proper time along the worldline. One can view $a(t)$ as being a conversion factor between a comoving coordinate distance and a physical distance. Two comoving objects separated by a proper distance $\ell_{\rm phys} = a(t)\ell_{\rm coord}$ recede from one another with a velocity

$$\frac{d\ell_{\rm phys}}{dt} = \dot{a}\ell_{\rm coord} = \frac{\dot{a}}{a}\,\ell_{\rm phys} \equiv H\ell_{\rm phys}, \tag{7.2}$$

where overdots denote time derivatives and

$$H = \frac{\dot{a}}{a} \tag{7.3}$$

is the Hubble parameter.

Homogeneity and isotropy imply that the energy–momentum tensor $T_{\mu\nu}$ can be expressed in terms of just two functions of $t$, the energy density $\rho(t)$ and the pressure $p(t)$. Einstein's equations then imply the Friedmann equation,

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi\rho}{3M_{\rm Pl}^2} - \frac{k}{a^2}, \tag{7.4}$$

where the Planck mass is related to Newton's constant by $M_{\rm Pl} = G_N^{-1/2} = 1.2 \times 10^{19}$ GeV.

The fact that $T_{\mu\nu}$ is covariantly conserved gives the equation

$$\dot{\rho} = 3H(\rho + p). \tag{7.5}$$

Given an equation of state, this determines the evolution of $\rho$ as the universe expands. The contents of the universe today can be classified into three components which, because their mutual interactions are relatively weak today, separately obey Eq. (7.5). Nonrelativistic matter, including both ordinary matter (baryons and electrons) and the dark matter, is essentially pressureless, and so obeys

$$\rho_{\rm matt} \sim a^{-3}, \tag{7.6}$$

which can be understood as conservation of particle number. Massless radiation (e.g., photons), with $p = \frac{1}{3}\rho$, obeys

$$\rho_{\rm rad} \sim a^{-4}. \tag{7.7}$$

Finally, current observations are consistent with the dark energy being a cosmological constant with $\rho_\Lambda = -p =$ constant.

Today, the dark energy dominates, with $\rho_\Lambda$ roughly three times $\rho_{\rm matt}$, and $\rho_{\rm rad}$ much smaller. These densities, together with the current value of $H$, determine the magnitude of the $k/a^2$ curvature term in the Friedmann equation. We can then work backward to trace the evolution of the universe at earlier times. Doing so, we see that while the dark energy makes the greatest contribution to $\rho$ today, nonrelativistic matter was dominant before that, and at earlier times (those in which we will be most interested here) the universe was in a radiation-dominated regime. We also find that the curvature term in the Friedmann equation, which makes only a small contribution today, was completely negligible at earlier times, so we can safely set $k = 0$ in our considerations. It then follows that $a \sim t^{2/3}$ during the matter-dominated era and $a \sim t^{1/2}$ during the radiation-dominated era.

The microwave background radiation today has a Planck spectrum corresponding to a temperature $T = 2.7$ K. Because the interactions of this radiation with matter (and with itself) are negligible today, this is only a nominal temperature, characterizing the spectrum of the microwave background, and not a measure of a system in thermal equilibrium.[1] However, at the higher densities of the early universe, the matter and radiation interacted rapidly enough to maintain the universe in true thermal equilibrium. The temperature fell as the universe expanded, but the expansion was slow and smooth enough that it can be treated as an adiabatic process, with entropy conserved and the entropy density $S$ obeying

$$a^3 S = \text{constant}. \tag{7.8}$$

In a radiation-dominated era the energy and entropy densities (in units with Boltzmann's constant equal to unity) are

$$\rho = \frac{\pi^2}{30}\mathcal{N}T^4, \tag{7.9}$$

$$S = \frac{2\pi^2}{45}\mathcal{N}T^3. \tag{7.10}$$

Here $\mathcal{N} = N_b + \frac{7}{8}N_f$, where $N_b$ and $N_f$ are the numbers of effectively massless bosonic and fermionic degrees of freedom. These count the number of spin states of particles with mass much less than $T$, and so are approximately stepwise constant, with a step downward each time the temperature falls below another particle mass.[2] Equations (7.8) and (7.10) imply that $aT$ is constant between such thresholds.

[1] Even in the absence of interactions, the redshifting of massless radiation is such that an initially thermal distribution maintains the Planck form, but with $T$ varying as $1/a$. This is the case here.

[2] Note that $\mathcal{N}$ was at least of the order of $10^2$ at early times; for temperatures above the electroweak scale the standard model particles alone give $\mathcal{N} = 106.75$.

Substituting Eq. (7.9) into Eq. (7.4) and using the fact that the expansion is adiabatic leads to a differential equation for $T(t)$ in the radiation-dominated era. Its solution is

$$T = \left(\frac{45}{16\pi^3}\right)^{1/4} \mathcal{N}^{-1/4} \sqrt{\frac{M_{\rm Pl}}{t}} = 0.55\,\mathcal{N}^{-1/4} \sqrt{\frac{M_{\rm Pl}}{t}}. \tag{7.11}$$

The integration constant here has been chosen so that $t = 0$ is the time at which the temperature and energy density diverge and the scale factor $a \to 0$. Of course, these are only formal statements, since the Friedmann–Robertson–Walker approximation must break down at sufficiently high $T$, and certainly cannot be trusted if $T$ is Planckian in size.

Causality considerations will be of particular importance for us. Consider a light signal emitted from $r = 0$ at time $t_0$. At a later time $t$, it will have traveled a coordinate distance (assuming a flat universe)

$$\ell_{\rm coord}(t_0, t) = \int_{t_0}^{t} \frac{dt'}{a(t')} \tag{7.12}$$

that corresponds to a physical distance

$$\ell_{\rm phys}(t_0, t) = a(t) \int_{t_0}^{t} \frac{dt'}{a(t')}. \tag{7.13}$$

Setting $t_0 = 0$ gives the size

$$d_H(t) = a(t) \int_{0}^{t} \frac{dt'}{a(t')} \tag{7.14}$$

of what is called the particle horizon. If two objects are separated by more than twice this horizon distance, their past light cones have no points in common, and the objects are causally disconnected. For a flat radiation-dominated universe, we find

$$d_H = 2t = 0.60 \mathcal{N}^{-1/2} \frac{M_{\rm Pl}}{T^2}. \tag{7.15}$$

An analogous calculation gives $d_H = 3t$ for a matter-dominated universe.[3]

## 7.2 Symmetry restoration and cosmological phase transitions

It is a common phenomenon that symmetries that are spontaneously broken at low temperature are restored at high temperature. The magnetization of a ferromagnet, which spontaneously breaks the rotational symmetry of the Hamiltonian,

[3] It is believed that there was an earlier era of cosmological inflation during which the universe, or at least our portion of it, expanded exponentially fast [120]. As a result of this, the actual horizon distances would be vastly larger than the expressions above; indeed, this is precisely how inflation explains the homogeneity of the presently observed portion of the universe. However, the expressions given here are the appropriate ones for determining the maximum causal influence of events occurring in post-inflationary times.

disappears above the Curie temperature. The crystal structure of a solid breaks both translational and rotational symmetry, but these are restored if the crystal is heated to its melting temperature.

A similar high-temperature symmetry restoration can occur in a quantum field theory [121–124]. To understand this, consider a weakly coupled theory in which a complex scalar field $\phi$ interacts with a massless fermion field $\psi$ and an Abelian gauge field $A_\mu$, giving them masses $G|\phi|$ and $g|\phi|$, respectively. At zero temperature the equilibrium value of $\phi$ is determined by minimizing the energy density of a uniform configuration, $V(\phi)$. [More precisely, one should find the minimum of the effective potential, $V_{\rm eff}(\phi)$, which includes the higher-order quantum corrections to the tree-level potential [125].] At a finite temperature $T$ the quantity to be minimized is the free energy density, usually expressed as a finite temperature effective potential $V_{\rm eff}(\phi, T)$.

With weak coupling, the various particle species can be treated as essentially ideal gases. The free energy density is then given, to a first approximation, by the sum of the zero-temperature energy density $V(\phi)$ and the ideal-gas free energies of the various particle species. The latter depend on the masses of the particles, which in turn depend on $\phi$. For $M \ll T$, the free energy density per spin degree of freedom of an ideal gas of bosons with mass $M$ is

$$F = -\frac{\pi^2}{90}T^4 + \frac{M^2}{24}T^2 + \cdots, \tag{7.16}$$

while for fermions of mass $M$ we have

$$F = -\frac{7\pi^2}{720}T^4 + \frac{M^2}{48}T^2 + \cdots. \tag{7.17}$$

Hence, if the tree-level potential is

$$V(\phi) = -\mu^2|\phi|^2 + \frac{1}{2}\lambda|\phi|^4, \tag{7.18}$$

with $\mu^2 > 0$, the finite temperature effective potential for small $|\phi|$ is [123]

$$V_{\rm eff}(\phi, T) = -\frac{\pi^2}{90}\mathcal{N}T^4 + (-\mu^2 + \sigma T^2)|\phi|^2 + O(|\phi|^3), \tag{7.19}$$

where

$$\sigma = \frac{1}{8}g^2 + \frac{1}{12}G^2 + \frac{1}{3}\lambda \tag{7.20}$$

reflects the contributions to the free energy from the $\phi$-dependent part of the masses of $A_\mu$, $\psi$, and $\phi$ itself, with the coefficients including factors for the various possible polarizations.

At zero temperature $\phi = 0$ is a local maximum of the potential and the symmetry is spontaneously broken. For $T > T^* = \sqrt{\mu^2/\sigma}$, the coefficient of $|\phi|^2$ is positive, and $\phi = 0$ is a local minimum of the effective potential. For sufficiently

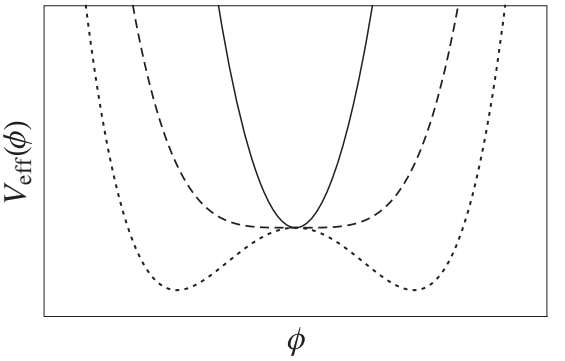


Fig. 7.1. Evolution of the shape of the finite temperature effective potential for a second-order phase transition. The curves correspond to $T > T_c$ (solid line), $T = T_c$ (dashed line), and $T < T_c$ (dotted line). Arbitrary constants have been added to make the curves coincide at $\phi = 0$.

large $T$ this is always the global minimum, but to determine whether this is the case for $T \sim T^*$ we need to know the behavior of the effective potential at all values of $\phi$. For weak coupling, this can be done by diagrammatic methods that sum all one loop vacuum graphs in the presence of a spatially uniform background $\phi$ [122]. The case of strong coupling is more difficult to address analytically, but can often be studied by numerical lattice field theory methods.

Generically, there are two possible behaviors, depending on whether or not the equilibrium value of $\langle\phi\rangle$ is continuous at the critical temperature. The former case, called a second-order transition, is illustrated in Fig. 7.1. In this example the origin goes directly from being the global minimum to being a local maximum at $T^*$, which is therefore the critical temperature $T_c$. As the universe cools below the critical temperature, $\langle\phi\rangle$ increases from zero until it reaches its zero-temperature value. If this is the case for the example of Eq. (7.19) and $g^2 \sim G^2 \sim \lambda$,

$$T_c = \sqrt{\frac{\mu^2}{\sigma}} \sim \sqrt{\frac{\mu^2}{\lambda}} \sim \langle\phi\rangle_0, \tag{7.21}$$

where $\langle\phi\rangle_0$ is the zero-temperature vacuum expectation value.

The other possibility, a first-order transition, is illustrated in Fig. 7.2. In this example, at very high temperature $\phi = 0$ is the only minimum of $V_{\text{eff}}$. An asymmetric local minimum appears at $T_1$, becomes degenerate with the symmetric minimum at $T_c < T_1$, and is the global minimum for $T < T_c$. In the example outlined above the symmetric minimum disappears at $T^*$, but for other theories it may persist down to $T = 0$. In a first-order transition the low-temperature phase does not emerge smoothly from the high-temperature phase. Instead, the transition proceeds by the nucleation of bubbles of the low-temperature phase, a process which we will examine further in Chap. 12. Assuming that the nucleation

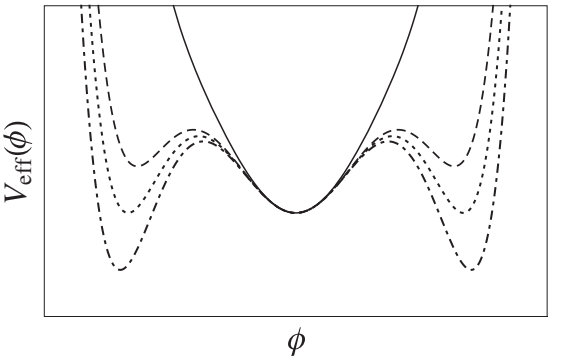


Fig. 7.2. Evolution of the shape of the finite temperature effective potential for a first-order phase transition. The curves correspond to $T \gg T_c$ (solid line), $T > T_c$ (dashed line), $T = T_c$ (dotted line), and $T < T_c$ (dot-dashed line). Arbitrary constants have been added to make the curves coincide at $\phi = 0$.

rate is large compared to the rate at which the universe is cooling, these bubbles expand and eventually merge to form a uniform low-temperature phase.[4]

In particular, it is believed that at $T \sim 10^2$ MeV the universe went from a QCD phase with manifest chiral symmetry and unconfined quarks, to the present confining phase with broken chiral symmetry. Earlier, at a higher temperature, $T_c \sim 10^2$ GeV, there was an electroweak transition from a phase in which the SU(2)×U(1) symmetry was manifest, to the current low-temperature phase in which this symmetry is spontaneously broken. If there is a grand unified theory, there would have been at least one, and possibly more, transitions corresponding to the breaking of the GUT symmetry at still earlier times.

## 7.3 The Kibble mechanism

Let us consider a phase transition from a symmetric phase characterized by a vanishing scalar field $\phi$ to an asymmetric phase in which the effective potential has degenerate minima at nonzero values of $\phi$. (For simplicity, I will refer to these as vacuum values of $\phi$, but it should be kept in mind that the minima of the effective potential at finite $T$ are not necessarily the same as those at $T = 0$.) If the transition is second-order, then as the universe cools past the critical temperature $\phi$ will become nonzero and move toward one of its vacuum values. Because the vacua are all physically equivalent, all vacuum values are equally probable. Although it would be energetically favorable for the same vacuum to

[4] If the nucleation rate is small relative to the cosmological expansion, the universe enters a regime of extreme supercooling, and the transition is never globally completed [126, 127]. For the present discussion I will assume that this is not the case, and that the transition is completed.

be chosen everywhere, the choice can only be uniform over a finite distance, leading to a system of domains characterized by a correlation length $\xi$.

If the transition is first order, then the vacuum can be uniform within a single bubble (at least until it collides with another bubble), but the choices in different bubbles will be uncorrelated. When the bubbles coalesce, a domain structure again appears, with the characteristic bubble size at coalescence playing the role of $\xi$.

Once the transition is completed, the dynamics will tend to smooth out the variations in the field at the domain boundaries. However, as pointed out by Kibble, there can be topological obstructions that prevent this, leading to the creation of topological defects [128].

Perhaps the simplest case to visualize is that with a discrete symmetry leading to two distinct vacua, with $\langle\phi\rangle = \pm v$. Once the fields have settled down after the transition there will be regions of positive $\langle\phi\rangle$ and ones with negative $\langle\phi\rangle$, with domain walls—(3+1)-dimensional generalizations of the kink—along the boundaries between them. Any region with volume more than a few times $\xi^3$ would be expected to have at least one domain wall traversing it.

A second possibility is that the phase transition corresponds to the breaking of a symmetry group $G$ to a subgroup $H$, with $\pi_1(G/H)$ being nontrivial. For definiteness, consider a theory where a complex scalar field $\phi$ develops a nonzero vacuum expectation value. Figure 7.3 shows a caricature of the domain structure along a two-dimensional spatial slice just after the transition is completed, with the arrows indicating the phase of $\phi$ in the various domains. The field dynamics will tend to align the phases of neighboring domains. However, this relaxation to a uniform phase cannot be complete, because there will inevitably be some domain junctions, such as the ones shown in Fig. 7.4, with a net vorticity that

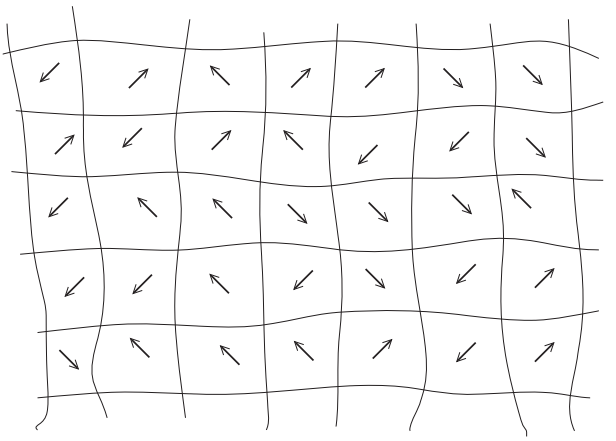

Fig. 7.3. Domain structure shortly after a phase transition in which a U(1) symmetry is broken by a nonzero complex scalar field. The arrows indicate the phase of the field.

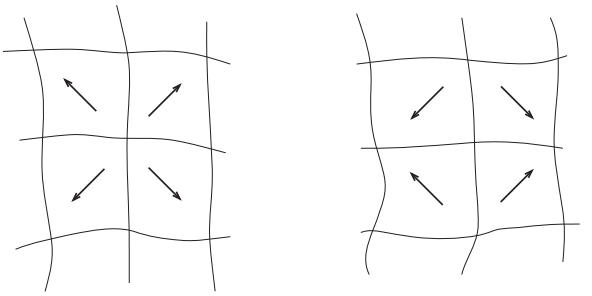

Fig. 7.4. Domain junctions leading to the formation of a vortex (left) or an antivortex (right).

cannot be smoothed away. Instead, these will lead to the formation of topological strings that appear as a vortices on this two-dimensional slice. The number of strings per unit area will be roughly

$$n_V \sim p_V \xi^{-2}, \tag{7.22}$$

where $p_V$ measures the probability of nontrivial vorticity arising at a domain junction from the random phases in the adjacent domains. The value of $p_V$ depends on the details of the theory, but it cannot be much less than order unity; a reasonable estimate is $p_V \sim 1/10$.

Finally, if the transition corresponds to a symmetry breaking with nontrivial $\pi_2(G/H)$, there can be point defects at domain junctions, leading to point-like solitons such as magnetic monopoles. By arguments analogous to those for the strings, we see that the initial density of these will be

$$n_M \sim p_M \xi^{-3}, \tag{7.23}$$

with $p_M$ again being a number not much less than unity.

The crucial quantity in these estimates is the correlation length $\xi$. Its value depends on the detailed dynamics of the fields involved in the transition and on the rate at which the universe cools past the critical temperature. Whatever the dynamics, we know that there cannot be causal correlations of the field on distances greater than the horizon, so that the initial domain size cannot be greater than $d_H$ at the time that the defects form [129]. One's first thought might be to take this to be the time when $T = T_c$, but this is not quite right. In a second-order transition there will still be thermal fluctuations back to the symmetric phase until the universe has cooled to the slightly lower Ginzburg temperature. In a first-order transition the defects will form at the time that the bubbles coalesce to complete the transition. Because of supercooling, this will be at a temperature somewhat less than the critical temperature. If we denote the temperature at which the defects form by $\tilde{T}_c$, then in the radiation-dominated era the causality bound on the initial correlation length is [129]

$$\xi \leq d_H(\tilde{T}_c) \sim \frac{M_{\rm Pl}}{\tilde{T}_c^2}. \tag{7.24}$$

This upper bound on $\xi$ implies lower bounds on the initial densities of the various topological defects. It must be stressed that in many, if not all, cases the actual value of $\xi$ will be much less than the horizon distance, so that these lower bounds may vastly underestimate the actual initial densities. They will, however, be sufficient for our purposes.

One might object to applying these arguments in a gauge theory, because they have been phrased in terms of the group orientation of the scalar field, which is not a gauge-invariant quantity. This is easily remedied. To give a concrete example, consider the production of monopoles in a theory where a triplet Higgs field breaks SU(2) to U(1). Now consider a spherical surface of radius $L \gg \xi$. According to Eq. (4.41), the topological charge contained within this surface is

$$N_\phi = \frac{1}{8\pi}\epsilon^{ijk} \int dS^i \, \hat{\boldsymbol{\phi}} \cdot \partial_j \hat{\boldsymbol{\phi}} \times \partial_k \hat{\boldsymbol{\phi}}. \tag{7.25}$$

Using Eq. (5.55), this can be rewritten as a surface integral with a gauge-invariant integrand,

$$N_\phi = \frac{1}{8\pi}\epsilon^{ijk} \int dS_i \left[\hat{\boldsymbol{\phi}} \cdot (\mathbf{D}_j \hat{\boldsymbol{\phi}} \times \mathbf{D}_k \hat{\boldsymbol{\phi}} - e\mathbf{F}_{jk})\right]. \tag{7.26}$$

We can analyze this integral by arguments similar to those we have been using. At any point on the integration surface either sign for the integrand is equally likely, and so we would expect the sign to be correlated only over distances of order $\xi$. Hence, the integral should be viewed as a sum over roughly $(L/\xi)^2$ patches, with signs assigned randomly in each patch. The difference in the numbers $N_+$ and $N_-$ of monopoles and antimonopoles should then be of the order of the square root of the number of patches,

$$|N_+ - N_-| \sim \frac{L}{\xi}. \tag{7.27}$$

Because our previous arguments showed that the total number of monopoles and antimonopoles was

$$N_+ + N_- \sim \left(\frac{L}{\xi}\right)^3, \tag{7.28}$$

we might have expected that $|N_+ - N_-| \sim (L/\xi)^{3/2}$. The fact that this quantity is only linear in $L$ indicates a correlation between the positions of the monopoles and the antimonopoles.

## 7.4 Gravitational and cosmological consequences of domain walls and strings

Both domain walls and cosmic strings would be recognized primarily by their gravitational effects. We will see that the effects of the walls are disastrous except at very low mass scales, thus placing very stringent conditions on theories with

spontaneous breaking of discrete symmetries. Strings, on the other hand, could quite plausibly be detected.

One would expect a planar domain wall to have large and obvious gravitational effects. In Newtonian gravity such a wall would give rise to an attractive force that was independent of distance. However, the general relativistic analysis leads to rather different results. Let us focus on a planar domain wall that is described by the extension of a one-dimensional scalar field kink solution to three dimensions. If the wall is in the $x$-$y$ plane, we have

$$\phi(x, y, z) = \phi_{\text{kink}}(z). \tag{7.29}$$

The energy–momentum tensor is

$$T_{\mu\nu} = \partial_\mu \phi \, \partial_\nu \phi - g_{\mu\nu} \mathcal{L}, \tag{7.30}$$

so that

$$\begin{aligned} T_{00} = -T_{11} = -T_{22} &= \frac{1}{2}\phi'(z)^2 + V(\phi(z)), \\ T_{33} &= \frac{1}{2}\phi'(z)^2 - V(\phi(z)), \end{aligned} \tag{7.31}$$

where the prime indicates a derivative with respect to $z$. Integrating across the thickness of the wall, we find negative pressures (i.e., positive tensions) in the $x$- and $y$-directions with magnitudes equal to the energy density per unit area, while the net pressure in the $z$-direction vanishes by virtue of the virial identity in Eq. (2.20). These negative pressures have a repulsive effect that is twice the attractive effect of the energy density, so that an observer at the wall sees test particles moving away from the wall with constant acceleration. However, the spacetime away from the wall is actually flat [assuming that $V(\phi)$ vanishes at its minima], very much like a higher-dimensional analogue of Rindler spacetime. In fact, one can find coordinates in which the metric on one side of the wall is that of a portion of Minkowski spacetime surrounded by a spherical wall that collapses to a minimum size and then expands, always with a constant outward acceleration [130–132].

More relevant for cosmology is the effect, not of a single wall, but of the network of domain walls produced via the Kibble mechanism during a phase transition where a discrete symmetry is broken. This network will evolve as the universe cools. The general tendency will be for walls to become more planar and for closed walls that enclose finite domains to contract until the domain has disappeared. The average size of the domains that remain will increase with time.

Even without a detailed study of the dynamics of this evolution, we can obtain a useful constraint just from causality. The same arguments that tell us that the correlation length just after the phase transition must be less than the horizon length at that time imply that the characteristic domain size at any later time cannot be greater than the horizon length at that time. It follows that at any

time $t$ the total domain wall area per unit volume must be at least of the order of $1/d_H(t)$. The energy density from domain walls is therefore

$$\rho_{\rm wall} \gtrsim \frac{\sigma}{d_H(t)}, \tag{7.32}$$

where $\sigma$ is the wall mass density per unit area; from our analysis of the kink solutions we know that $\sigma \sim m^3/\lambda$, where $m$ and $\lambda$ are the mass and coupling constant associated with the fields underlying the domain wall.

In both the radiation- and matter-dominated regimes the horizon distance is proportional to $t$, implying that the wall energy density falls more slowly than those of radiation and matter, and could potentially come to dominate them. However, the subsequent evolution of a wall-dominated universe is sufficiently different as to be clearly in conflict with cosmological observations. Excluding this possibility places an upper limit on $\sigma$ [133]. For example, requiring that horizon-crossing walls not dominate the energy density today implies that [48, 134]

$$\sigma \lesssim (100\,{\rm MeV})^3. \tag{7.33}$$

A stronger bound, $\sigma \lesssim (1\,{\rm MeV})^3$, is obtained by considering the effects of walls on the anisotropy of the cosmic microwave background [48, 133].

Hence, any domain walls surviving to the present must be associated with very low-energy physics that has so far escaped discovery. Domain walls with a higher energy scale could have existed in the past, but only if they later ceased to be stable. This instability could result from a later phase transition that changed the vacuum structure of the theory, or it could happen if the discrete symmetry whose breaking led to the domain wall was only approximate. In the latter case domain walls would persist while the energy difference between the vacua on either side was negligible compared to the cosmic temperature, but at sufficiently low temperature the pressure from the lower-energy true vacuum would cause the regions of higher-energy vacuum to shrink and the walls to disappear.

More detailed discussions of the evolution of networks of cosmic domain walls and of their observational consequences can be found in [48, 134].

The situation with strings is rather different. By analogy with the argument for domain walls, a minimum expectation is that there should be at least one string crossing the visible universe. For strings arising in gauge theories, where the energy density is concentrated in a narrow core, the gravitational effects are hardly as dramatic as that of a domain wall. For a straight solitonic string only one component of the tension is nonzero, with the same magnitude as the energy density, so a nearby test particle feels neither attraction or repulsion. There is, however, a conical singularity at the string. This has a lensing effect, so that one signal of such a string would be double images of galaxies located behind it.

A major focus has been on the study of the network of strings that would emerge from a suitable phase transition. It was suggested that with an appropriate choice of the symmetry-breaking scale these could have served as the seeds

for the density inhomogeneities that grew and evolved to the structure that we see in the universe today. As a result, considerable efforts, both analytic and numerical, have been devoted to the study of the problem. However, it has now become clear that while strings can lead to inhomogeneities of roughly the right magnitude, they cannot reproduce the detailed features of the cosmic microwave background spectrum. These are instead much better fit by inhomogeneities arising from slow-roll inflation, although the possibility of a small contribution from strings is not ruled out. These considerations place an upper bound on the energy per unit length $\mu$ of the string. Some recent studies [135, 136] quote bounds of roughly

$$G\mu \lesssim 7 \times 10^{-7}, \tag{7.34}$$

corresponding to a symmetry-breaking scale no higher than $10^{15}$ GeV or so. Bounds of a roughly similar range are obtained by considerations of the effects of the gravitational radiation from the strings.

A comprehensive discussion of cosmic strings is given in [48]. Some more recent reviews are [137, 138].

## 7.5 Evolution of the primordial monopole abundance

Magnetic monopoles are produced by the Kibble mechanism with an initial density given by Eq. (7.23). Assuming weak gauge coupling, the monopole mass is greater than the critical temperature of the transition where they are formed. They are therefore nonrelativistic, and their initial abundance is considerably greater than what it would be in thermal equilibrium. Monopoles disappear through monopole–antimonopole annihilation although, as we will see, the dilution of the monopole density by the cosmic expansion eventually brings an end to this annihilation.

Monopole–antimonopole annihilation in the early universe was studied by Zeldovich and Khlopov [139] and, with a particular emphasis on GUT monopoles, by Preskill [140]. It is perhaps best viewed as a two-step process in which the monopole and antimonopole are first captured into a Coulomb bound state, and then subsequently move down to lower bound states and eventually annihilate. It is the capture process that limits the annihilation rate. This is a purely electromagnetic process, and so does not depend on the details of the monopole's non-Abelian core.

The essential requirement for capture is that the initially free monopole and antimonopole, each with mass $m$ and with magnetic charges $\pm Q_M$, lose enough of their initial kinetic energies that they can form a bound state. At high temperatures they are moving in a plasma of relativistic charged particles. They undergo Brownian motion with a mean free path

$$\ell \sim \frac{1}{CT}\sqrt{\frac{m}{T}}, \tag{7.35}$$

where $C \sim (1-5)\mathcal{N}_c$ if the number of charged degrees of freedom is in the range $1 \lesssim \mathcal{N}_c \lesssim 100$. As long as this is less than the capture radius,

$$r_c \sim \frac{Q_M^2}{4\pi T}, \tag{7.36}$$

where the negative Coulomb potential energy of the pair becomes comparable to their thermal kinetic energy, the drag forces exerted by the plasma can dissipate enough energy for the pair to be captured. However, once the universe has cooled below the temperature $T_1 \sim (4\pi)^2 m/(C^2 Q_M^4)$ where $\ell = r_c$, capture is only possible if an initially unbound monopole–antimonopole pair loses enough energy through bremsstrahlung to become bound.

In either temperature regime the time derivative of the monopole density $n_M$ can be written as

$$\dot{n}_M = -D n_M^2 - 3\frac{\dot{a}}{a} n_M, \tag{7.37}$$

where $D$ represents the effects of the annihilation processes and the last term is the effect of the cosmic expansion.

In the high-temperature regime the Coulomb attraction felt by a monopole at a distance $r$ from an antimonopole is opposed by the drag forces from the plasma, with the net effect being a drift velocity

$$v_{\rm Drift} \sim \frac{Q_M^2}{4\pi} \frac{1}{CT^2 r^2} \tag{7.38}$$

toward the antimonopole. If the typical monopole separation is $d \sim n_M^{-1/3}$, the capture time is

$$\tau \sim \frac{d}{v_{\rm Drift}} \sim \frac{4\pi}{Q_M^2} \frac{CT^2}{n_M} \tag{7.39}$$

and

$$D \sim \frac{1}{\tau n_M} \sim \frac{Q_M^2}{4\pi} \frac{1}{CT^2}. \tag{7.40}$$

In the low-temperature regime, with an initial monopole thermal velocity $\sim \sqrt{T/m}$, the cross-section for radiative capture via bremsstrahlung emission is

$$\sigma_{\rm rad} \sim \left(\frac{Q_M^2}{4\pi T}\right)^2 \left(\frac{T}{m}\right)^{3/5}, \tag{7.41}$$

giving

$$D \sim v\sigma_{\rm rad} \sim \left(\frac{Q_M^2}{4\pi m}\right)^2 \left(\frac{m}{T}\right)^{9/10}. \tag{7.42}$$

With the above two expressions for $D$, we can now solve Eq. (7.37) to obtain the evolution of the monopole density. However, it is better to separate the effects of annihilation and expansion by working instead with the monopole-to-entropy ratio

$$r = \frac{n_M}{S}. \tag{7.43}$$

If the expansion is adiabatic, Eq. (7.8) implies that

$$\dot{r} = -DSr^2. \tag{7.44}$$

It is convenient to express $r$ as a function of temperature rather than time. In a radiation-dominated regime with the number of massless degrees of freedom constant, so that $\dot{T}/T = -\dot{a}/a$,

$$\frac{dr}{dT} = -\frac{1}{\dot{T}}\, DSr^2 = \left(\frac{\pi\mathcal{N}}{45}\right)^{1/2} DM_{\rm Pl}\, r^2. \tag{7.45}$$

Integrating this gives

$$r(T) = \left[\frac{1}{r_{\rm init}} + \frac{1}{r_*(T)}\right]^{-1}, \tag{7.46}$$

where $r_{\rm init}$ is the initial monopole to entropy ratio and

$$r_*(T) = \left[\left(\frac{\pi\mathcal{N}}{45}\right)^{1/2} M_{\rm Pl} \int_T^{T_c} dT'\, D(T')\right]^{-1}. \tag{7.47}$$

When the high- and low-temperature expressions for $D(T)$ are substituted into Eq. (7.47), the integration divides into two regimes, both dominated by the region near $T_1$. Thus at temperatures below $T_1$

$$r_* \approx \left(\frac{45}{\pi\mathcal{N}}\right)^{1/2} \left[C\left(\frac{Q_M^2}{4\pi}\right)^3 + 10C^{-1/5}\left(\frac{Q_M^2}{4\pi}\right)^{9/5}\right]^{-1} \frac{m}{M_{\rm Pl}}. \tag{7.48}$$

For $C \approx 10^2$, $\mathcal{N} \approx 10^2$, and $Q_M = 2\pi/e$, and with

$$m_{17} \equiv \frac{m}{10^{17}\,{\rm GeV}}, \tag{7.49}$$

this gives

$$r_* \approx 10^{-10}\, m_{17}. \tag{7.50}$$

Under the same assumptions Eqs. (7.23) and (7.24) give a lower bound

$$r_{\rm init} \gtrsim p\, \frac{1}{S(\tilde{T}_c)\,[d_H(\tilde{T}_c)]^3} \approx \left(\frac{\tilde{T}_c}{M_{\rm Pl}}\right)^3. \tag{7.51}$$

If $\tilde{T}_c$ is about an order of magnitude smaller than the monopole mass, this gives

$$r_{\rm init} \gtrsim 10^{-9}\, m_{17}^3. \tag{7.52}$$

From Eq. (7.46), we see that at large times $r$ will be given by the lesser of Eqs. (7.50) and (7.52).

## 7.6 Observational bounds and the primordial monopole problem

Let us now compare the predictions for the monopole-to-entropy ratio from the previous section to the various observational bounds on the current monopole density. The simplest of these is obtained by noting that $r$ is related to the monopole fraction of the critical density, $\Omega_M$, by

$$r \approx 10^{-27}\, m_{17}^{-1} \Omega_M. \tag{7.53}$$

Even in the rather implausible case that monopoles were to account for all of the mass usually attributed to dark matter, so that $\Omega_M \approx .25$, this exceeds both $r_*$ and $r_{\rm init}$ unless $m \lesssim 10^{12}$ GeV.

Other density bounds follow from the limits on the monopole flux $F$ in our galaxy. If the monopoles are uniformly distributed throughout the universe, $n_M = F/v$, where $v$ is the typical monopole velocity. If instead they cluster with the galaxies, the average value of $n_M$ would be up to five orders of magnitude smaller for a given value of $F$.

If there were no galactic magnetic field, monopoles in the galaxy would have typical velocities on the order of $10^{-3}c$, which is both the virial velocity in the galaxy and its peculiar velocity with respect to the rest frame of the cosmic microwave background. However, our galaxy does have a magnetic field, with a magnitude of approximately $3\times 10^{-6}$ gauss, that is coherent over distances of the order of $10^{23}$ cm. A monopole with magnetic charge $2\pi/e$ would be accelerated by this field to a velocity

$$v_{\rm mag} \sim \begin{cases} c, & m \lesssim 10^{11}\ \text{GeV}, \\ 10^{-3} m_{17}^{-1/2} c, & m \gtrsim 10^{11}\ \text{GeV}. \end{cases} \tag{7.54}$$

Hence, monopoles with masses less than about $10^{17}$ GeV will be accelerated sufficiently to be ejected from the galaxy, and thus certainly do not cluster with our galaxy.

The acceleration of these monopoles drains energy from the galactic field. Requiring that the rate of this loss be small compared to the time scale on which the field can be regenerated (roughly $10^8$ yrs) gives the Parker bound [141, 142]

$$F < F_{\rm Parker} = \begin{cases} 10^{-15}\,\text{cm}^{-2}\text{sr}^{-1}\text{sec}^{-1}, & m \lesssim 10^{17}\text{GeV}, \\ 10^{-15} m_{17}\,\text{cm}^{-2}\text{sr}^{-1}\text{sec}^{-1}\,, & m \gtrsim 10^{17}\text{GeV}. \end{cases} \tag{7.55}$$

The two cases here reflect the fact that while the lighter monopoles are carried along the magnetic field lines, the heavier ones experience only small deflections by the field. Applying similar reasoning to an earlier seed field from which the present galactic field developed gives the somewhat stronger bound [143]

$$F < \left[m_{17} + (3\times 10^{-6})\right] 10^{-16}\,\text{cm}^{-2}\text{sr}^{-1}\text{sec}^{-1}. \tag{7.56}$$

Reasoning along these lines can also be applied to the magnetic fields in galactic clusters, giving a bound which, although less certain, is about two orders of magnitude tighter than the Parker bound [144].

There are also limits from direct searches for monopoles in cosmic rays. For monopoles with $v > 10^{-4}c$, the MACRO experiment [145] places an upper bound of about $10^{-16}\,\mathrm{cm}^{-2}\mathrm{sr}^{-1}\mathrm{sec}^{-1}$. Somewhat stronger bounds have been obtained by other experiments, but these are limited to monopoles with higher velocities.

Even more stringent bounds apply for GUT monopoles that catalyze baryon number violation via the Callan–Rubakov effect. The essential idea is that such monopoles would be captured by compact astrophysical objects. They would then catalyze baryon decay, with the energy released in the decay leading to an increase in the luminosity of the object. A variety of bounds have been obtained by considering neutron stars [146–150], white dwarfs [151], and Jovian planets [152]. These depend on the details of the astrophysical scenario, such as whether monopoles captured by a progenitor star survive its collapse to a white dwarf or neutron star, and on the degree to which monopole–antimonopole annihilation reduces the accumulated density in the object. The bounds obtained in this manner lie in the range

$$F\left(\frac{\sigma_{\Delta B} v}{10^{-27}\,\mathrm{cm}^2}\right) \lesssim (10^{-18} - 10^{-25})\,\mathrm{cm}^{-2}\mathrm{sr}^{-1}\mathrm{sec}^{-1}, \tag{7.57}$$

where $\sigma_{\Delta B}$ is the cross-section for catalysis of baryon number violation.

For a GUT monopole mass of $10^{17}$ GeV, with the monopoles not clustering with the galaxies, we have upper bounds on $r$ of $10^{-26}$ from both the mass density and Parker bounds, $10^{-27}$ from direct observation, and perhaps as low as $10^{-36}$ for monopoles that catalyze baryon number violation. These range from 15 to 25 orders of magnitude below the predictions of Eqs. (7.50) and (7.52). Even taking into account the uncertainties in the various estimates involved, there is a very clear conflict between the cosmological predictions and the observational bounds. This conflict persists for monopole masses down to about $10^{12}$ GeV and to even lower masses if the monopoles catalyze baryon number violation.

This poses a serious problem for any grand unified theory. All such theories necessarily predict the existence of superheavy monopoles, and any plausible unification scale predicts that at least one species of these has a mass well above $10^{12}$ GeV. [Lighter multiply-charged monopoles, such as the SO(10) example discussed in Sec. 6.3.2, could have masses this low.] One might therefore decide to simply abandon all such theories. However, the idea of unification is sufficiently attractive as to motivate attempts to find a resolution of this primordial monopole problem that is consistent with grand unification.

The most attractive solution to this problem is based on the inflationary universe scenario, in which the universe undergoes a period of exponential expansion followed by a reheating process in which vacuum energy is converted to particles with a thermal distribution. (Indeed, it was consideration of the primordial

monopole problem that led Guth to the idea of inflation [120].) If the inflation takes place after monopoles have been produced, any pre-existing monopoles will be diluted by an exponential factor. As long as the reheating after inflation does not raise the temperature of the universe above the critical temperature of the GUT phase transition, the present-day value of $r$ will be unobservably small.

Although inflation is the most widely accepted solution, there is an alternative proposal that is of interest, if only for illustrative purposes, that was put forth by Langacker and Pi [153]. Consider, for example, a scenario with the following series of phase transitions:[5]

$$\mathrm{SU}(5) \to \mathrm{SU}(3) \times \mathrm{SU}(2) \times \mathrm{U}(1) \to \mathrm{SU}(3) \to \mathrm{SU}(3) \times \mathrm{U}(1). \tag{7.58}$$

Monopoles are formed at the first transition, when an SU(5) adjoint field $\phi$ gets a nonzero vacuum expectation value, because $\pi_2[\mathrm{SU}(5)/\mathrm{SU}(3) \times \mathrm{SU}(2) \times \mathrm{U}(1)]$ is nontrivial. At the next transition, the breaking of the U(1) symmetry leads to the formation of strings. The monopoles cannot survive as free objects after this transition, because $\pi_2[\mathrm{SU}(5)/\mathrm{SU}(3)] = 0$. Instead, they become bound to strings that have a monopole and an antimonopole at opposite ends. Because the energy of a string is proportional to its length, the monopole–antimonopole pair are drawn to each other by a constant force, leading to a rapid and efficient annihilation.

Free monopoles can exist again once the U(1) symmetry is restored. One might therefore expect that the horizon bound together with the Kibble argument would produce monopoles with a minimum density of roughly one per horizon volume. (Since the critical temperature for this transition must be of the order of the electroweak scale, this would not conflict with observation.) This reasoning is incorrect. When the fields settle down after the first transition, the only constraint on $\phi(x)$ is that it be continuous. This allows the configuration of $\phi$ on a sphere with radius $\gg \xi$ to correspond to any element of $\pi_2[\mathrm{SU}(5)/\mathrm{SU}(3) \times \mathrm{SU}(2) \times \mathrm{U}(1)]$. After the second transition, when an additional field $\psi$ also becomes nonzero, the requirement is that both $\phi$ and $\psi$ be continuous. If there are constraints on the relative orientation of $\phi$ and $\psi$, these may eliminate the $\phi$ configurations corresponding to nontrivial elements of $\pi_2[\mathrm{SU}(5)/\mathrm{SU}(3) \times \mathrm{SU}(2) \times \mathrm{U}(1)]$, so that monopoles do not reappear when $\psi$ becomes zero again.

[5] The final transition may seem a bit odd, since the low-temperature phase has higher symmetry than the high-temperature one. This is possible in a theory with several scalar fields. Recall that the effect of finite temperature is to add an effective scalar field mass $\sigma T^2$. For the example considered in Sec. 7.2, $\sigma$ was given in Eq. (7.20). In order that $V(\phi)$ be bounded from below, $\lambda$, and hence the right-hand side of Eq. (7.20), must be positive. In the more general case, with arbitrary numbers of scalar, spinor, and gauge fields, the contributions to $\sigma$ from the gauge and Yukawa couplings remain positive. However, it is possible to have scalar quartic couplings of both signs that give a net negative contribution to $\sigma$ while still keeping $V(\phi)$ bounded from below. With a negative $\sigma$, this unconventional ordering of symmetries is possible.

To put this more formally [154], let us suppose that we have a theory with two fields such that $\langle\phi\rangle$ by itself breaks $G$ to $H$, while the combined effect of $\langle\phi\rangle$ and $\langle\psi\rangle$ is to break $G$ to a subgroup $\hat{H} \subset H$. Every configuration of $\phi$ corresponds to an element of $\pi_2(G/H)$, which is assumed to be nontrivial. Every combined configuration of $\phi$ and $\psi$ corresponds to an element of $\pi_2(G/\hat{H})$. Considering only the $\phi$ in such a combined configuration gives an element of $\pi_2(G/H)$, thus giving a map from $\pi_2(G/\hat{H})$ into $\pi_2(G/H)$. However, this map need not be onto, and it could even be the case that all elements $\pi_2(G/\hat{H})$ map to the identity element of $\pi_2(G/H)$.

If we start with a symmetric phase with $\langle\phi\rangle = \langle\psi\rangle = 0$ and then go directly to one with $\langle\phi\rangle \neq 0$, $\langle\psi\rangle = 0$, and unbroken symmetry $H$, then configurations corresponding to all elements of $\pi_2(G/H)$ can be created. On the other hand, if the breaking is from $G$ to $\hat{H}$ to $H$, then in the final state only those elements of $\pi_2(G/H)$ that lie in the image of $\pi_2(G/\hat{H})$ can arise; if this image is just the identity element of $\pi_2(G/H)$, then no monopoles are created by the Kibble mechanism.