Part I

ELEMENTARY PARTICLES IN COSMOLOGY

A. Dolgov

ELEMENTARY PARTICLES AND COSMOLOGY.

A. D. Dolgov ITEP, Moscow 117259, Russia

Lectures presented at VII Brazilian School of Cosmology and Gravity Rio de Janeiro, 1-14 August, 1993

Chapter 1

Introduction.

A magnificent progress in our understanding of Nature at the two extremes of very small and very large scales that was reached in the last 10-20 years is reflected in the name "Standard Model" used for both the elementary particle theory and for cosmology. They are, of course, two different standard models, though closely related to each other, and their symbiosis has proved to be one of the reasons for the success. It was not the main reason for the success in particle theory but definitely played a major role in the development of the modern cosmology. Now the table is turned to some extent and cosmology gives a strong indication to an incompleteness of the standard model in particle physics which may lead to a further progress in the latter. At the present day all experiments in high energy physics are in a very good agreement with theoretical calculations and only astronomy spoils the picture by presenting the data which demands new physics, the so-called physics beyond the standard model. One may argue that there is not much of success if it is necessary to invoke new physics for an explanation for some phenomena in cosmology. Still there is a big difference between the old and the new cosmology because earlier a lot of cosmological problems could not be resolved with any reasonable physical hypothesis and now their solution is achieved with relatively mild physical assumptions. I do not mean, of course, that all the problems are solved. On the opposite, it seems that we see only an edge of a very interesting new world and may hope to see more in the nearest future.

In this lectures I will consider these two standard models of particle physics and cosmology and analyze their possible inconsistencies. As a rather separate sub-

ject I will discuss the cosmological machinery which permits to obtain information of the particle properties supplementary to that accessible from the direct laboratory measurements. This will be done on the example of neutrinos. We will start with a brief introduction into the theory of elementary particles based on the famous symmetry group $SU(3) \times SU(2) \times U(1)$ with the minimum particle content and with the discussion of possibilities as well as theoretical need for extensions of the minimal model. After that the standard cosmological scenario with inflation, baryo- and nucleo- synthesis and a short discussion of the large scale structure formation will be presented. An attention will be given to the kinetics of elementary particles in the hot primeval plasma and to the phase transition in the theories with spontaneously broken symmetries. Related to that is the discussion of the properties of the topological defects and their cosmological role which will be shortly presented. I also plan to consider the problems of the cosmological constant and the problem of the initial state in the big-bang theory as well as a possibility of the modification of the standard cosmological scenario in this connection.

Chapter 2

The Minimal Standard Model (MSM) in Particle Physics.

A. Classification.

The Minimal Standard Model is now very well known to the community. So I will be rather brief here. The fundamental matter constituents are presented by two very much different (at least at low energies) sets of particles: quarks (q) and leptons (l). They all are fermions having spin 1/2. Inside each set the particles are grouped into pairs or doublets which are also called families.

Known leptons form three doublets: electronic (ν_e, e^-), muonic (ν_μ, μ^-) and taonic (ν_τ, τ^-). Electric charge of the first member of the family, neutrino, is equal to zero and the second member is charged with the charge equal to (-1) (in proton charge units). Charged leptons are relatively light particles (τ maybe an exception) with the masses: $m_e = 0.5$ MeV, $m_\mu = 105$ MeV, and $m_\tau = 1.8$ GeV. All direct measurements of the neutrino masses are compatible with zero. The corresponding upper bounds are (see Particle Data Group).

$$m_{\nu_e} < 8eV \tag{0.1}$$

$$m_{\nu_\mu} < 500KeV \tag{0.2}$$

$$m_{\nu_T} < 32MeV \tag{0.3}$$

There are known also (almost) three quark families: (u,d), (c,s), and (t,b). Electric charge of the first member of each quark family is $(+2/3)$ and that of the second member is $(-1/3)$. Of these 6 quarks only one, the heaviest t, has not yet been observed experimentally. The expected value of the mass found from radiative corrections to the observed processes is around 150 GeV. Recently the CDF group at Fermilab announced the observation of several events which may be considered as an evidence of the observation of t-quark with the mass around 175 GeV. The masses of the other quarks are: $m_d = 4$ MeV, $m_u = 7$ MeV, $m_s = 150$ MeV, $m_c = 1.2$ GeV, and $m_b = 4.9$ GeV. These results are valid for masses measured at large momentum transfer when the effects of the quark interactions are not essential. By this reason the masses are called the masses of current quarks (in contrast to the constituent quarks). The effect of interactions is especially important for the light quarks.

Each quark may exist in three completely degenerate states which have the same mass and exactly the same interactions and are called the color states. So quarks are described by color (they are indistinguishable in color) and flavor. The latter refers to the different quark families.

At this point we need 12 parameters (masses) to describe the model. The value of the masses is one of the unsolved mysteries of the standard model. They span the range from practically or exactly zero up to $O(100)$ GeV. At the present day we have very little knowledge of how these values can be understood. The most puzzling are probably the small masses of neutrinos. Though only the ν_e mass is strongly bounded by experiment, cosmology gives similar upper limits for ν_μ and ν_τ if they are stable (see below). With our present knowledge we cannot exclude that all or some neutrino masses are exactly zero. There are some data however implying nonvanishing values for neutrino masses, namely the deficit of the solar neutrinos and the anomaly in the fluxes of muonic and electronic neutrinos. These phenomena may be explained by neutrino oscillations which are possible and moreover natural if the masses are nonzero. Cosmology also hints that the mass of one (or all) of the neutrino species may be in the range of 10 eV. Massive neutrinos are natural candidates for the so called hot dark matter.

B. Gauge principle and interactions.

Interactions of the fundamental fermions is described by the exchange of vector and scalar bosons. There is also gravitational interaction induced by the exchange of tensor bosons (gravitons) but we neglect them in this section. The strongest part of the interaction associated with the exchange of vector particles

is rather well understood theoretically and is described by the principle of gauge invariance. We will first discuss this principle for the simplest case of electromagnetic interactions for which it was historically first formulated. The operator (or the wave function) ψ describing an electrically charged particle always enters the Lagrangian in the combination $\psi^*\psi$. This combination is invariant with respect to the phase rotation $\psi \to \exp(ie\theta)\psi$ where e is the electric charge of the field ψ. For a constant θ the kinetic part of the Lagrangian which contains $\psi^*\partial\psi$ (for fermions) or $(\partial\psi^*)(\partial\psi)$ (for bosons) is also invariant. However it is rather unnatural to make the phase rotation all over the Universe simultaneously with the same phase. Much more natural is the phase which depends on the space-time point, $\theta = \theta(x^\mu)$. In this case however the kinetic term is not invariant and transforms in accordance with $\delta(\partial_\mu\psi) = ie\partial_\mu\theta\psi$. To compensate this variation and to make the theory invariant with respect to phase transformation with an arbitrary $\theta(x)$ one has to introduce a new vector field $A_\mu(x)$ which appears in the combination

$$\partial_\mu \to \partial_\mu - ieA_\mu \tag{0.4}$$

Under the phase rotation (or in other words under gauge transformation) the vector field A_μ should transform as

$$A_\mu \to A_\mu - i\partial_\mu\theta \tag{0.5}$$

To make the field A_μ a dynamical variable one has to introduce a kinetic term for it which must be invariant with respect to the gauge transformation. It dictates the form of the latter to be $\mathcal{L} = -F_{\mu\nu}^2/4$ with $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. We got in this way the usual Maxwell electrodynamics.

For more complicated theories the gauge transformations include not only phase rotation but may also transform one field into another. So the transformation is not a multiplication by a number but by a matrix in the field space. The idea of generalization of gauge transformations of quantum electrodynamics which form $U(1)$-group to the case of more complicated groups belongs to Yang and Mills (1954).

The transformations of $U(1)$ commutes with each other i.e. two transformations made in different order lead to the same result. Such groups are called Abelian groups. Theories based on non-Abelian groups are much more interesting and rich. The simplest example of non-Abelian group is the group of rotations when e.g. rotations around axes X and Y lead to a different result than the same rotations made in inverse order.

Let us consider a theory invariant with respect to a non-Abelian compact continuous group g. Let ψ_j $(j = 1, 2, ...n)$ is an n-component field transforming by an irreducible representation of the group. This means that under transformations of g the components of ψ are expressed through each other and so that they can not be divided into separate pieces that do not mix. The action of g on ψ can be described by $n \times n$ matrix G_k^j:

$$\psi_k' = G_k^j \psi_j \tag{0.6}$$

The matrix of any transformation from g can be written as

$$G = exp(i \sum_{k=1}^{N} \alpha_k J_k) \tag{0.7}$$

where α_k are some numbers which are called the parameters of transformation and matrices J_k which can be taken Hermitian are the group generators. The number of independent group generators is called the group order and the number of the generators which commute with each other is the group rank. For example the group of rotations is the group of the third order and the first rank.

If ψ is a spinor field the kinetic term in the Lagrangian has the form

$$L_k = \frac{i}{2} \overline{\psi}^j \gamma^\mu \overleftrightarrow{\partial}_\mu \psi_j \tag{0.8}$$

where $\overline{\psi} \gamma^\mu \overleftrightarrow{\partial}_\mu \psi = \overline{\psi} \gamma^\mu \partial_\mu \psi - (\partial_\mu \overline{\psi}) \gamma^\mu \psi$, γ_μ are the Dirac matrices, and $\overline{\psi}$ belongs to the conjugate to ψ representation so that the product $\overline{\psi}^j \psi_j$ is invariant under the group transformations:

$$\overline{\psi}^j \psi_j \to \overline{\psi}^j exp(-i\alpha^k J_k) exp(i\alpha^l J_l) \psi = \overline{\psi} \psi \tag{0.9}$$

If however the group transformations are local i.e. the parameters α^k are functions of coordinates then expression (0.8) is of course noninvariant. There appear terms proportional to $\partial_\mu \alpha^k$. To compensate them one has to substitute covariant derivatives instead of the usual ones as it has been done in the considered above case of electrodynamics:

$$\partial_\mu \psi \to D_\mu \psi = \partial_\mu \psi + igA_\mu \psi \tag{0.10}$$

Now A_μ is not a single field but the matrix:

$$A_\mu = \sum_k^N A_\mu^k(x) J_k \tag{0.11}$$

One can check up that substitution (0.10) makes the term $\overline{\psi}\gamma^\mu D_\mu \psi$ invariant under local transformations of group g if A_μ is transformed as

$$A_\mu \to A_\mu + \frac{1}{g}\partial_\mu \alpha^k J_k - i\alpha^k [J_k, A_\mu] \tag{0.12}$$

where $[J_k, A_\mu] = J_k A_\mu - A_\mu J_k$ is the commutator of J_k and A_μ.

The kinetic term of the fields A_μ^k can be formally written as above $(-1/2)Tr(F_{\mu\nu}^2)$ but now $F_{\mu\nu}$ has the form

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + ig[A_\mu, A_\nu] \tag{0.13}$$

where the extra term is the commutator of the matrices A_μ and A_ν. In the Abelian case this term evidently vanishes. The product $TrF_{\mu\nu}^2$ is invariant under transformation (0.12). Tr means trace of the product of matrices J_k. The coefficient $1/2$, instead of $1/4$ in the Abelian case, is connected with the usual normalization condition $TrJ_k = 1/2$.

Let us formulate the main lessons that we have learned. First, if a symmetry is realized locally then there should be present vector fields which mediate interactions between particles. Second, if the local symmetry is more complicated than simple $U(1)$ there appear not a single vector field, as in electrodynamics, but several ones the number of which is equal to the order of the group. Third, the condition of gauge invariance fixes the interaction of matter fields with the vector fields which is of the form $g\overline{\psi}A_\mu\gamma^\mu\psi$. Fourth, nonabelian vector fields in contrast to photons directly interact with each other. This is connected with the commutator term $ig[A_\mu, A_\nu]$ in eq.(0.13). In that sense one can say that the vector fields themselves are charged. Radiation or absorption of nonabelian vector fields change internal quantum numbers or in other words charges of the source.

There exists a deep but of course not complete analogy between the theory of nonabelian gauge fields and the theory of gravity. The source of gravitational

field is the energy of a system. But gravitons themselves possess energy. This leads to direct interactions between gravitons and correspondingly to nonlinear theory. Selfinteraction is inherent to gravitational field as well as to nonabelian gauge fields.

Gauge field theories possess a very important property of renormalizability. Field theory deals with systems with infinite number of degrees of freedom. Because of that quantizing of the theory leads to infinite values of some quantities. We know from quantum mechanics that the energy of the ground state of oscillator is nonvanishing, $E_0 = \omega/2$. Quantum field can be represented as an infinite set of oscillators with eigenvalues ranging from 0 to infinity. The ground state of such a system has of course infinite energy. This energy is called the energy of zero oscillations of vacuum. We return to this phenomenon later in connection with the problem of cosmological constant and now only note that in quantum field theory this difficulty is simply avoided by shifting the scale of energy by (infinite) value E_0. This is an example of the simplest renormalization.

There can appear other infinite quantities. For example the perturbation theory corrections to mass or charge of a particle are expressed through divergent integrals. This is also connected with contribution of infinite number of degrees of freedom. If such infinities appear in a finite number of physical quantities one can still work with the theory assuming that these quantities are free parameters of the theory which are to be determined from experiment. This is the essence of renormalization procedure.

For formal manipulations with divergent quantities in the theory one or other regularization procedure is used which consists in cutting off the number of degrees of freedom. As a rule this cutoff effectively means introduction of a maximum momentum p_{max}. In renormalizable theories dependence on p_{max} disappear after the renormalization is made.

Regularization is an essential ingredient of any quantum field theory. For gauge theories the usual condition imposed on regularization is that it should keep the symmetry of unregularized theory. In some cases however symmetries which peacefully coexist at the level of classical theory can not do that with quantum corrections taken into account. The reason for that is that it is impossible to make a regularization which respects all the symmetries of initial theory. This phenomenon of breaking classical conservation laws by quantum corrections is called quantum anomaly.

The existence of quantum anomalies unexpectedly has resulted in very interesting cosmological consequences. In particular the so called axial anomaly leads to nonconservation of baryonic charge. The hypothesis about existence of the new light particle - axion is also closely connected with this anomaly. Because of anomaly in the trace of energy momentum tensor the massless particle production by conformally flat gravitational field becomes possible. We discuss these phenomena below in some more detail.

C. Spontaneous symmetry breaking.

Nonabelian gauge fields were not popular in elementary particle physics for a long time because it seemed necessary that vector bosons should be massless and no other massless vector bosons except for photon were known. The situation has changed when the idea of spontaneous symmetry breaking came to particle physics. One says that a symmetry is broken spontaneously if the Lagrangian is symmetric but the ground state is not. Ground state in quantum field theory is called vacuum. The symmetry manifests itself in existence of several (possibly infinitely many) ground states which have equal energy and connected by symmetry transformations. It is said in this case that vacuum is degenerate. A well known example of a system with spontaneously broken symmetry is ferromagnetic in which the symmetry with respect to rotations is lost.

The main features of spontaneous symmetry breaking in field theory can be understood in the following simple example of a complex scalar field with the Lagrangian

$$L = (\partial_\mu \phi^*)(\partial^\mu \phi) - V(|\phi|) \tag{0.14}$$

This Lagrangian is symmetric with respect to phase rotations $\phi(x) \to \phi(x)exp(i\Phi)$ where Φ does not depend on x. For a free (noninteracting) scalar field $V(|\phi|) = m^2|\phi|^2$. In our example we chose $V(|\phi|)$ as

$$V(|\phi|) = \lambda(|\phi|^2\eta^2)^2 \tag{0.15}$$

The parameter η is real. One can easily see that the state $\phi = 0$ is unstable. Indeed the classical equation of motion of ϕ is

$$\partial_\mu \partial^\mu \phi + 2\lambda(\phi^*\phi - \eta^2)\phi = 0 \tag{0.16}$$

It follows from this equation that long wave fluctuations of ϕ near $\phi = 0$ exponentially rise with time:

$$\phi \sim exp[(2\lambda\phi_0^2 - k^2)^{1/2}t]\exp(ikx) \tag{0.17}$$

Here k is the wave vector of the fluctuation.

The state $\phi = \eta$ is stable. To demonstrate that let us introduce instead of complex field ϕ two new real fields through the equation

$$\phi = [\eta + \phi_1(x)]\exp[i\alpha(x)/\phi_0] \tag{0.18}$$

The Lagrangian rewritten in terms of fields α and ϕ_1 has the form

$$L = (\partial_\mu\alpha)(\partial^\mu\alpha)(1 + \frac{\phi_1}{\eta})^2 + (\partial_\mu\phi_1)(\partial^\mu\phi_1) - \lambda(2\eta + \phi_1)^2\phi_1^2 \tag{0.19}$$

Small fluctuations of $\phi_1(x)$ and $\alpha(x)$ are evidently stable. Their Fourier modes has the usual form $\exp(-i\omega t + ikx)$. For the field ϕ_1 the dispersion relation is $\omega^2 = k^2 + 4\lambda\eta^2$. So the mass of the quanta of ϕ_1 is $2\eta\lambda^{1/2}$. For the α-field $\omega^2 = k^2$ and so this field is massless. The vanishing of the mass of α is connected with the possibility of movement on the bottom of the potential $V(|\phi|)$ without changing the energy. Thus the system has infinitely degenerate set of ground states, $\phi = \eta\exp(i\Phi)$ with arbitrary phase Φ. Once the ground state is chosen the symmetry breaks. This is the spontaneous symmetry breaking.

The appearance of a massless field in this example is not accidental but a consequence of the general theorem by Goldstone (1961). According to this theorem massless bosons always appear when a global symmetry is broken.

The word global is very essential. If the symmetry is realized locally, that is the transformations with variable phase $\Phi(x)$ are permitted, the situation is quite different. As we know the Lagrangian should contain in this case the vector field which compensates the variation induced by the coordinate dependence of the phase:

$$L = -\frac{1}{4}F_{\mu\nu}^2 + |D_\mu\phi|^2 - \lambda(|\phi|^2 - \eta^2)^2 \tag{0.20}$$

The state $\phi = 0$ is as before unstable and $\phi = \eta$ is stable. The field $\alpha(x)$ introduced through eq.(0.18) now does not correspond to a physical degree of freedom because it can be removed by the appropriate choice of the phase $\Phi(x)$. The field $\phi(x) = \eta + \phi_1(x)$ can be made real. Such a choice of the phase is called unitary gauge. In this gauge the Lagrangian (0.20) has the form

$$L = -\frac{1}{4}F_{\mu\nu}^2 + (\partial_\mu\phi)^2 + g^2A_\mu^2(\eta^2 + 2\eta\phi_1 + \phi_1^2) - \lambda(2\eta\phi_1 + \phi_1^2)^2 \qquad (0.21)$$

This Lagrangian describes massive scalar field ϕ_1 with the mass $2\eta\lambda^{1/2}$ and massive vector field with the mass $2^{1/2}g\eta$. The latter is given by the term $g^2A^2\eta^2$.

Thus starting from the theory of massless gauge field interacting with complex scalar field ϕ we came to the theory of massive vector and scalar fields. The extra degree of freedom associated with $Im\phi$ forms the longitudinal state of the vector field (massless vector field has only transverse components). In other words the vector field A_μ gets the mass because of interaction with the vacuum condensate of the scalar field $< \phi >= \eta$. This phenomenon was discovered in papers by Higgs(1964), Englert and Brout (1964), and Guralnik, Hagen and Kibble (1964). This scalar field is now called Higgs field. Analogous appearance of the photon mass in superconductors was found by Ginsburg and Landau (1950).

D. Electroweak theory.

The theory of electroweak interactions is constructed essentially along the same lines with the only difference that the symmetry group is slightly more complicated so that the processes with charge transition like

$$e^- + p \to \nu + n$$

can be described. Here the charge of lepton rises by one unit, $(e^- \to \nu)$, and the charge of nucleon respectively goes down, $(p \to n)$. The intermediate boson responsible for this process should be charged and so should interact with photons. We have already seen that gauge bosons interact with each other if the symmetry group is nonabelian.

The simplest after $U(1)$ group is $SU(2)$. This is the group of rotations in which double valued (spinor) representations are permitted which corresponds to half integer values of angular momentum. And it happens that this simplest possible nonabelian group indeed describes electroweak interactions. The group $SU(2)$ has

three generators which on the spinor representation can be chosen as the linear combination of the Pauli matrices

$$\tau_+ = (\tau_1 + i\tau_2)/2^{1/2} = \begin{pmatrix} 0 & 2^{1/2} \\ 0 & 0 \end{pmatrix} \tag{0.22}$$

$$\tau_- = (\tau_1 - i\tau_2)/2^{1/2} = \begin{pmatrix} 0 & 0 \\ 2^{1/2} & 0 \end{pmatrix} \tag{0.23}$$

$$\tau_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{0.24}$$

Correspondingly this group demands introduction of three gauge bosons W^+, W^-, and A^0. Tau-matrices determine the form of interaction of these bosons with quarks and leptons:

$$L_{int} = g\overline{\psi}\tau_i A^i_\mu \gamma^\mu \psi = g\overline{\psi}(\tau_+ W^-_\mu + \tau_- W^+_\mu + \tau_3 A^0_\mu)\gamma^\mu \psi \tag{0.25}$$

where ψ is a two-component spinor in the group space, e.g. (ν, e)

The form of matrix τ_3 shows that if the field A^0 is identified with photon then the charges of up and down components of the spinor are equal by their, size and opposite by the signs. We know that this is not true however. By this reason the choice of $SU(2)$ as a gauge group is not satisfactory if quarks and leptons are indeed spinors. To solve the problem the slightly more complicated group $SU(2) \times U(1)$ was proposed. In other words an extra vector field B^0 was added which interacts like the photon with the hypercharge prescribed to each particle. The value of the hypercharge was fixed by the condition

$$Q = I_3 + Y/2 \tag{0.26}$$

where Q is the electric charge of a particle, I_3 is the third component of its weak isospin, and Y is the hypercharge. Evidently $Y = -1$ for leptonic dublets and $Y = -1/3$ for quarks. The photon is a linear combination of the fields A_0 and B_0:

$$\gamma = B\cos\theta_W + A\sin\theta_W \tag{0.27}$$

which proves to be massless. The orthogonal combination is the intermediate vector boson of weak interactions

$$Z_0 = -A\cos\theta_W + B\sin\theta_W \tag{0.28}$$

The angle θ_W is called the weak angle or Weinberg angle.

It is noteworthy that the condition that the massless vector boson interacts just with the electric charge of particles is fulfilled in this model automatically. This is ensured by the conservation of electromagnetic $U(1)$-invariance in models with single dublet Higgs field. The models with several Higgs multiplets should be organized in such a way so that to maintain electromagnetic $U(1)$ which is not always natural.

As we have seen gauge bosons can become massive as a result of spontaneous symmetry breaking when condensate of Higgs field ϕ is formed. To realize that there should exist a scalar field interacting both with isotriplet (W^+, W^-, A^0) and with isosinglet B^0. This means that ϕ should have both nonzero hypercharge and weak isospin. The simplest possible choice of ϕ as $SU(2)$-dublet has proved to be just the right one because it gives the ratio of W and Z masses in excellent agreement with experiment.

Besides the vector fields the model predicts the existence of the neutral scalar field with the mass $2^{1/2}\lambda^{1/2}\eta$. The other three components of the complex dublet ϕ turn into longitudinal components of massive W and Z bosons.

Interaction of gauge fields with quarks and leptons is described in the standard way by changing the usual derivatives to the covariant ones (see eq. (0.10). In comparison to electrodynamics however the theory is slightly more complicated. The point is that only left-handed neutrinos take part in weak interactions. The latter are obtained by the projection operator

$$\psi_L = \frac{1+\gamma_5}{2}\psi \tag{0.29}$$

For massless particles this projector picks out definite helicity states, namely negative ones for neutrinos and positive ones for antineutrinos. (Recall that helicity is the projection of the particle spin on the direction of its momentum.)

Evidently two-component neutrino and four-component electron cannot be members of the same dublet. The only way out is the assumption that the weak dublets consist only of left-handed particles. We know however that right-handed

electron also exists and interacts with photons because it is charged. Since the right-handed neutrino, if exists, does not take part in weak interactions there is no partner for the right-handed electron and we have to assume that its isospin is zero but hypercharge is not, $Y(e_R) = -2$. This gives the right value of the electric charge.

Thus the right-handed electron interacts only with B^0 and the interaction strength is twice as large as that of e_L. The right-handed quarks are described analogously. The difference in interactions of right-handed and left-handed particles leads to the famous phenomenon of parity nonconservation in weak interactions.

The separation on left-handed and right-handed particles has invariant meaning for massless particles only. Thus in the phase where symmetry is unbroken, $< \phi >= 0$, we have to assume that quarks and leptons are massless. They acquire mass as a result of spontaneous symmetry breaking if there exists the Yukawa type interaction:

$$f(\overline{\psi}_L \psi_R \phi + \overline{\psi}_R \psi_L \phi^+) \tag{0.30}$$

For nonzero $< \phi >= \eta$ and e.g. for $\psi_{L,R} = e_{L,R}$ this gives for the electron mass $m_e = f_e \eta$. The same mechanism can ensure nonzero masses of other leptons and quarks with appropriately chosen coupling constants f.

The Higgs bosons generically should interact with a different combination of fermions than gauge fields. Hence the frame in which the quark mass matrix is diagonal does not coincide with the frame where the matrix of W and Z interactions is diagonal. In particular W-boson transforms u-quark into $d \cos\theta_c + s \sin\theta_c$ where θ_c is the Cabibbo angle. The account of the third quark generation makes the mixing matrix more complicated. It is parametrized by three angles and is called Kobayashi - Maskawa matrix. Because of these mixings weak interactions do not conserve strangeness, charm, etc. Analogous phenomenon is not observed in the lepton sector. Electronic, muonic, and taonic charges are conserved with a rather good accuracy. This conservation is an evidence in favor of vanishing neutrino mass.

The model is now complete. In this example the main features inherent to the gauge models with spontaneously broken symmetry are well seen. First the theory is formulated for massless particles with possible exception of scalar ones. Masses appear as a result of formation of the Higgs field (or fields) condensate. This approach does not permit to make any prediction about the fermion masses. They

are determined by arbitrary Yukawa coupling constants f_j changing from 10^{-6} to 10^{-1} for electron and b-quark respectively. Any natural explanation of such hierarchy is absent now and it makes the Higgs boson construction subject to criticism. On the other hand it presents the only known way to construct a consistent renormalizable theory of weak interactions and it is just the contribution of the scalar particles which stops the catastrophic rise of vector boson interactions with energy.

The basic principles of this theory were formulated in the works by Glashow, Salam, and Weinberg about 30 years ago. The intermediate W and Z bosons were discovered soon after that with the masses predicted by the theory. To complete the picture one has to find the Higgs bosons. Their searches however is much more difficult not only because of their very weak interactions but also because their masses in contrast to W and Z are not known in advance. The search of the Higgs bosons is one of the central problems in experimental high energy physics .

Note in conclusion that though the considered model is called the unified model of electroweak interactions this is not exactly true. The constants g_2 and g_1 corresponding respectively to the weak isospin group $SU(2)$ and to the weak hypercharge group $U(1)$ are independent. In that sense a real unification of electromagnetic and weak interactions is not achieved. This has led to numerous attempts to construct models in which the $SU(2) \times U(1)$ group is a subgroup of a larger group which is characterized by a single coupling constant. These are the so called models of grand unifications.

E. Qunatum Chromodynamics.

The principle of local gauge invariance is also successfully applied to the theory of strong interactions. The gauge group in this case is $SU(3)$ acting in the space of quark colors. Each quark is a color triplet forming a fundamental representation of $SU(3)$. The eight generators of $SU(3)$ in fundamental representation can be chosen as the Gell-Mann matrices $\lambda_k/2 (k = 1, 2...8)$ which are essentially the Pauli matrices with an added "empty" first, second, and third column and row for nondiagonal ones. As for the diagonal matrices they are $\lambda_3 = diag(1, -1, 0)$ and $\lambda_8 = diag(1, 1, -2)/\sqrt{6}$. The gauge bosons connected with the color $SU(3)$ are called gluons. The gluon matrix A_μ is defined in accordance with eq. (0.11) where $J_k = \lambda_k/2$ and the field strength $G_{\mu\nu}$ is given by eq. (0.13).

In a compact form the Lagrangian of quantum chromodynamics can be written as follows

$$L = -\frac{1}{2} Tr G_{\mu\nu} G^{\mu\nu} + \overline{q}(i\hat{D} - M)q \tag{0.31}$$

where $q = [q_{if}]$ is a colon in the space of quark colors i and flavors f. The covariant derivatives do not depend on the quark flavors but through the matrix A_μ depends on color:

$$(\hat{D})^{f'i'}_{fi} = \gamma^\mu [\delta^{i'}_i \delta^{f'}_f \partial_\mu + ig\delta^{f'}_f (A_\mu)^{i'}_i] \tag{0.32}$$

The mass matrix on the opposite does not depend on color but depends on quark flavor

$$M = \delta^{i'}_i M^{f'}_f \tag{0.33}$$

By unitary transformation of the quark basis the mass matrix can be diagonalized

$$M = diag(m_u, m_d, m_c, m_s, ...) \tag{0.34}$$

Note that the transformation diagonalizing M keeps the covariant derivative matrix $\hat{D}$ diagonal.

If the coupling constant g is known the theory is complete. There is a subtlety however in determination of g. It is not inherent to QCD only but exists in any quantum field theory. It is especially eminent in QCD because of relatively large interaction strength. The point is the following. Coupling constants in quantum field theory are not constant but change with the momentum transfer or in other words with the distance between particles. Thus if one says about the numerical value of g it makes sense only if the value of the momentum transfer q is specified. The physical mechanism producing the dependence $g(p^2)$ is the existence of virtual particles around the source characterized by the charge g. The role of these particles depends on the distance from the source. Let us consider for example the emission of gluon by quark. The effective coupling constant $g(p^2)$ is the sum of the bare coupling constant g_0 and the corrections connected with the virtual fermionic and gluonic pairs. The virtual fermions lead to the charge screening i.e. to the decrease of charge with rising distance. This phenomenon is long known in quantum electrodynamics and is absolutely evident physically. It remains in nonabelian theories. A new thing connected with selfinteractions of gluons is the charge antiscreening produced by virtual gluonic pairs. The mechanism of antiscreening may be understood

(in a simplified way) as the difference in signs of electric (Coulomb) and magnetic interactions: the same sign charges repulse but the same direction currents attract.

A more convenient parameter in the field theory is not the charge g but the square of it, $\alpha = g^2/4\pi$. In electrodynamics $\alpha_{EM} = e^2/4\pi \approx 1/137$ is called the fine structure constant. It can be shown that α depends on the momentum transfer in the limit of large p^2 as

$$\alpha^{-1}(p_1^2) - \alpha^{-1}(p_2^2) = \frac{b}{2\pi} \ln \frac{p_1^2}{p_2^2} \tag{0.35}$$

where b is a constant coefficient. This formula is obtained by summing up the perturbation theory expansion in small parameter α. It is valid when p is large in comparison with the virtual particle masses. If $p \leq m$ $\alpha(p)$ tends to a constant. The coefficient b depends on the gauge group G and on the number of fermion multiplets. For $G = SU(N_c)$ with N_f fundamental fermions

$$b = \frac{11}{3} N_c - \frac{2}{3} N_f \tag{0.36}$$

The first term here originates from the loops with gauge bosons while the second comes from fermionic ones. In particular for quantum chromodynamics with six quark flavors ($N_c = 3, N_f = 6$) $b = 7 > 0$. In electrodynamics photonic selfinteraction is absent and $b = -\frac{2N_f}{3} < 0$. The change of sign of b in nonabelian gauge theories gives rise to the new and very important phenomenon of asymptotic freedom. In pure electrodynamics the charge increases with rising momentum and in chromodynamics it goes down. The larger is the energy of particles the weaker is the interaction. Though it was discovered in the sixties (Vanyashin and Terentyev, 1965; Khriplovich, 1969) its importance was understood only after the papers by Politzer (1973) and Gross and Wilczek (1973) and has a tremendous impact on the further development of the theory.

Thanks to asymptotic freedom there is a hope to get a theoretical description of the primeval plasma in the early Universe possibly up to the Planck temperatures and densities, $\rho = m_{Pl}^4 \approx 10^{95} g/cm^3$, $T = m_{Pl} \approx 10^{32} K$.

With the decreasing momentum transfer or with the increasing distance the interaction between colored objects becomes very strong, the perturbation theory in α breaks and the structure of the theory drastically changes. At small distances

quarks can be considered as free point-like particles. With the increasing distance such simple description becomes wrong. The colored field strength rises up and prevents from separation of free quark or gluon. It is assumed that gluonic field lines form a narrow tube in contrast to the spread Coulomb field and this results in the potential linearly rising with the distance. The transition from the phase of free quarks and gluons to the phase where confinement effects are essential takes place at the distance of the order of 10^{-13}cm or $p = 100$Mev. This qualitative picture is in a good agreement with experiment but at the moment no consistent derivation of it from the first principles is known.

Quantum chromodynamics predicts that hadronic vacuum has a very complicated structure. In particular there is a condensate of quark-antiquark pairs in vacuum (Gell-Mann, Oakes, and Renner, 1968):

$$< \bar{q}q >_{vac} \approx -(250MeV)^3 \tag{0.37}$$

This value is derived from the the masses of pseudoscalar mesons (π, K, η). The point is that the QCD Lagrangian, with light quarks only, possesses the approximate chiral $SU(3)_L \times SU(3)_R$-symmetry which corresponds to separate $SU(3)$-transformation of left-handed $q_L = (1 + \gamma_5)q/2$ and right-handed $q_R = (1 + \gamma_5)q/2$ quarks. The quark masses break the symmetry but since they are small in comparison with characteristic hadronic energy scale the symmetry should manifest itself in the spectrum of hadrons giving rise to particles with close masses and opposite parities. This is not observed in experiment however. The absence of degeneracy in parity could be explained if the symmetry $SU(3)_L \times SU(3)_R$ is broken spontaneously so that only $SU(3)_{L+R}$ is left. The latter is the usual color symmetry of left-handed and right-handed quarks jointly. As we have seen the spontaneous symmetry breaking leads to massless Goldstone bosons. Such particles indeed exist. These are pseudoscalar mesons and especially the lightest among them π-meson. Its mass is not exactly zero but still much smaller than masses of other hadrons. The deviation of m_π from zero is connected with the approximate nature of the chiral symmetry which in turn is connected with nonzero quark masses. It can be shown that m_π^2 is proportional to $m_q < \bar{q}q >$. Such arguments permit to express the value of quark condensate through the known masses of pseudoscalar mesons.

It can be shown that gluonic condensate is also nonvanishing in QCD vacuum (Vainshtein, Zakharov, and Shifman, 1978,1979):

$$< G^2 >_{vac} \approx (800MeV)^4 \tag{0.38}$$

This numerical value is obtained from the vector meson masses. We recollect these condensates when we discuss the problem of the cosmological constant.

Spontaneous breaking of the chiral symmetry at least in part is connected with specific vacuum fluctuations of gauge fields called instantons. These fluctuations are very interesting by itself independently of the chiral symmetry. As we know the Lagrangian of gauge fields is $L = -G_{\mu\nu}^2/2$ where $G_{\mu\nu}$ is expressed through the potentials A_μ by eq. (0.13). The evident extremum of the action

$$S = \int L\, d^4x \tag{0.39}$$

is $F_{\mu\nu} = 0$. The potential A_μ is not necessary zero but could be the so called purely gauge potential:

$$A_\mu(x) = -\frac{i}{g} U(x) \partial_\mu U^{-1}(x) \tag{0.40}$$

where $U(x)$ is unitary unimodular $N \times N$ matrix. It can be checked that for this $A_\mu(x)$ the field strength tensor $F_{\mu\nu}$ vanishes. We assume that the matrices $U(x)$ are nonsingular, do not depend on time and that $U(x) \to 1$ when $x \to \infty$. In this case all infinite points are equivalent and the total three dimensional space is equivalent to three dimensional sphere S^3. The matrices U realize continuous mapping of S^3 on the gauge group G. Each mapping belongs to a particular class determined by the number of windings of S^3 on G. We consider first a simpler case of mapping of one dimensional sphere S^1 on G. S^1 is topologically equivalent to straight line with identified points $+\infty$ and $-\infty$. As an example of mapping let us take

$$V_q(x) = \exp\left[iq\pi \frac{xf(x)}{\sqrt{x^2+\rho^2}}\right], q = 0, \pm 1, \pm 2, ... \tag{0.41}$$

where $f(x)$ is an arbitrary function.of x satisfying condition $f(x \to \infty) \to 1$. Evidently $V_q \to 1$ when $x \to 1$. The transformation V_q realizes q-fold winding of S_1 on $SU(1)$. The transformations corresponding to different values of q can not be continuously transformed into each other.

The representatives of different classes corresponding to $SU(2)$-group can be chosen analogously:

$$U_q(x) = \exp\left[i\pi q\left(1 + \mathbf{x}\tau\frac{f(x)}{\sqrt{x^2+\rho^2}}\right)\right], q = 0, \pm 1, \pm 2, \ldots \qquad (0.42)$$

The case of arbitrary G can be reduced to this one because mapping of S_3 on G can be represented as mapping on its $SU(2)$-subgroup. The factor q characterizing different classes is called topological charge.

According to that, the potentials $A_\mu(x)$ corresponding to the vacuum state that is to $G_{\mu\nu} = 0$ are also divided into separate classes which cannot be continuously transformed into each other in the manifold with zero strength $G_{\mu\nu}$. If the transformations with $G_{\mu\nu} \neq 0$ are permitted then it proves possible to continuously deform the potential from one class to that in another class. This means that in infinitely dimensional space of field variables $A_\mu(x)$ the potential energy of gluonic field has a set of degenerate minima corresponding to pure gauge potentials. These minima are separated by barriers where $G_{\mu\nu} \neq 0$. It can be shown that this potential is periodic.

If these barriers were impenetrable the periodic structure of the potential would be unessential. Interactions and mixing between the states with different topological charges would be absent. In reality the barriers are penetrable. The probability of tunneling through the barrier can be evaluated in the same way as in quantum mechanics in quasiclassical approximation. It is determined by the factor $\exp(-S)$ where S is the action calculated on the trajectory connected the states which are separated by the barrier. The tunneling is permitted if the action is finite. The action reaches minimum on the solutions of the classical equations of motion in imaginary time, $t \to it$. Analogous approach is valid also in the field theory. The solution describing the transition from the state with $q = 0$ at $t = -\infty$ to the state with $q = 1$ at $t = +\infty$ is called instanton. The solution corresponding to the transition with $\Delta q = -1$ is called antiinstanton. These solutions were discovered by Belavin, Polyakov, Schwarts, and Tyupkin (1975).

The action for the Yang-Mills fields can be written as

$$S = -\frac{1}{2}\int TrG_{\mu\nu}^2\, d^4x = -\frac{1}{2}\int Tr[\pm G_{\mu\nu}\tilde{G}_{\mu\nu} + \frac{1}{2}(G_{\mu\nu} \pm \tilde{G}_{\mu\nu})^2]\, d^4x \qquad (0.43)$$

where $\tilde{G}_{\mu\nu} = \frac{1}{2}\varepsilon_{\mu\nu\alpha\beta}G_{\alpha\beta}$ and $\varepsilon_{\alpha\beta\mu\nu}$ is the antisymmetric in all four indices tensor, $\varepsilon_{1234} = +1$.

The quantity $G\tilde{G}$ is the divergence of the four-vector

$$G\tilde{G} = \partial_\mu K_\mu \tag{0.44}$$

$$K_\mu = 2\varepsilon_{\mu\nu\alpha\beta}(A_\nu \partial_\alpha A_\beta + \frac{2i}{3} g A_\nu A_\alpha A_\beta) \tag{0.45}$$

By the Gauss theorem the total derivative can be transformed into the integral over infinity and usually vanishes because the fields sufficiently fast disappear at infinity. This is not true however in the case we are considering. The action is finite if the field strength $G_{\mu\nu}$ quickly decreases but this is not obligatory for the potentials. Thus the integral

$$\int G\tilde{G}\, d^4x = \int \partial_\mu K_\mu \, d^4x \tag{0.46}$$

may be nonvanishing. This expression of course disappear for purely gauge potentials. It can be presented as a difference of two integrals of K_t over three-dimensional hypersurfaces $t = +\infty$ and $t = -\infty$:

$$Tr \int K_t \, d^3x = 16\pi^2 q/g^2 \tag{0.47}$$

We see from this expression that K_t is the density of topological charge of vacuum fluctuations.

It is easy to see that for fixed $\int G\tilde{G}\, d^4x$ the action S possesses minimum in imaginary time at

$$G_{\mu\nu} = \pm \tilde{G}_{\mu\nu} \tag{0.48}$$

The solutions of this equation are called (anti)instantons. It can be checked up that $G_{\mu\nu}$ satisfying eq. (0.48) simultaneously satisfy the equation of motion $D_\mu G_{\mu\nu} = 0$. The explicit solution of eq. (0.48) can be written as

$$A_\mu = \frac{2}{g}\eta_{\mu\nu}\frac{(x - x_0)_\nu}{(x - x_0)^2 + \rho^2} \tag{0.49}$$

$$G_{\mu\nu} = -\frac{4}{g}\eta_{\mu\nu}\frac{\rho^2}{[(x - x_0)^2 + \rho^2]^2} \tag{0.50}$$

where $\eta_{\mu\nu} = \eta_{\mu\nu i}\sigma_i/2, (i = 1,2,3)$, σ_i are the Pauli matrices, and $\eta_{\mu\nu i}$ are the t'Hooft symbols:

$$\eta_{\mu\nu i} = -\eta_{\nu\mu i} = \begin{cases} \varepsilon_{\mu\nu i}, & \text{if } \mu,\nu = 1,2,3 \\ \delta_{\mu i}, & \text{if } \nu = 4. \end{cases} \tag{0.51}$$

The following equation is valid for instantons

$$Tr \int G\tilde{G}\, d^4x = 16\pi^2/g^2 \tag{0.52}$$

This means that instantons indeed describe the transitions from the vacuum state with topological charge q at $t = -\infty$ to the vacuum state with topological charge $(q+1)$ at $t = +\infty$. The instanton action is

$$S_i = -8\pi^2/g^2 \tag{0.53}$$

Since the action is finite the vacuum in quantum chromodynamics should be a superposition of gauge potentials with different topological charges. The wave function of the vacuum is of the form

$$\Psi(A) = \sum_{q=-\infty}^{+\infty} C_q \psi_q(A) \tag{0.54}$$

where the functionals ψ_q are concentrated on the fields with topological charge q. The coefficients C_q are determined from the condition that under a gauge transformation which changes q the wave functional $\Psi(A)$ acquire only a phase factor

$$\Psi(A) \to \Psi(A) exp(i\theta) \tag{0.55}$$

where θ is an arbitrary constant phase. To ensure this the coefficients C_q should have the form

$$C_q = exp(iq\theta) \tag{0.56}$$

Note that wave functional (0.54) resembles the Bloch representation of the electron wave function in a periodic potential with θ being the quasimomentum.

Possible vacuum states are characterized by different values of θ which are strictly conserved but what value is realized in our world is not known. Consideration of a state with definite value of θ is equivalent to the addition to the Lagrangian the term

$$\mathcal{L}_\theta = \frac{g^2}{16\pi^2}\theta Tr G\tilde{G} \tag{0.57}$$

As it has been already mentioned this expression is a total derivative and gives in the action $S = \int L_\theta d^4x$ the topological charge of the vacuum fluctuations.

Expression (0.57) is not invariant with respect to space reflections. It is a pseudoscalar in contrast to the scalar G^2. So the addition of L_θ leads to parity nonconservation in strong interactions. It is known from experiment however that effects of parity violation are very weak and correspondingly the bound $\theta < 10^{-9}$ is valid. The attempts to find a natural mechanism of such a small θ (Peccei and Quinn, 1977) have led to the hypothesis of the existence of a new very light pseudoscalar particle - axion (Weinberg, 1978; Wilczek, 1978). We will discuss this hypothesis below and now only note that the axion, if exists, would lead to very interesting cosmological and astrophysical consequences.

F. More about quantum anomaly.

The terms of the same type as given by eq. (0.57 can appear because of parity breaking in the quark mass matrix:

$$L_m = \bar{q}(m_1 + im_2\gamma_5)q \tag{0.58}$$

The term proportional to $m_2\gamma_5$ which breaks parity conservation can be removed by the transformation

$$q \to \exp(i\alpha\gamma_5)q \tag{0.59}$$

If $\alpha = -(1/2)atan(m_2/m_1)$ then

$$L_m \to \sqrt{m_1^2 + m_2^2}\bar{q}q. \tag{0.60}$$

The kinetic term $\bar{q}\hat{D}q$ is not changed under this transformation. It seems that we have exiled the parity violation from the theory but we have forgotten about the

chiral anomaly. In the theory with massless quarks the classical action is invariant not only under the transformation $q \to exp(i\alpha)q$ but also under transformation (0.59). The usual machinery of the field theory leads in this case to the conservation of the vector and axial currents $\bar{q}\gamma_\mu q$ and $\bar{q}\gamma_\mu\gamma_5 q$. However the conservation of this two currents is mutually compatible in classical theory only. For example the amplitude of the transition of the axial current into two vector bosons through triangle of virtual fermions is given by a divergent expression which can be regularized in such a way that either vector or axial current is conserved but not both. We know from experiment that conserved is the vector current and this dictates the choice of the regularization. To this end one can use the Pauli-Villars method. The latter consists of introducing fictitious regulator fields q_r with masses m_r tending to infinity. After the renormalization is done the results should not depend on m_r. The contribution of the diagram with virtual q_r should be subtracted out the corresponding diagram with physical quarks. Such regularization evidently respects the vector current conservation but spoils the conservation of the axial current because $m_r \neq 0$ and so $\partial_\mu(\bar{q}_r\gamma_\mu\gamma_5 q_r) = 2m_r\bar{q}_r\gamma_5 q_r$. Hence the account of regulators in the triangle diagram of leads to the nonconservation of the axial current (Adler, 1969; Bell, Jackiw, 1969):

$$\partial_\mu J_\mu = \frac{g^2\beta}{16\pi^2} TrF\tilde{F} \tag{0.61}$$

where β is a numerical coefficient determined by the type and number of the virtual particles forming the triangle in the considered Feynman diagram. For one quark loop $\beta = 1$.

It is essential that the anomaly leads to the nonconservation of the left currents $J_\mu^L = (J_\mu^V + J_\mu^A)/2$ to which the intermediate bosons of the weak interactions are coupled. This can be shown to break the renormalizability of the theory. The disaster is cured by the quark-lepton symmetry. Due to it the contributions of quarks and leptons into triangle diagram are canceled out and the anomaly does not arise in the physical currents.

The reader might has got the impression that the axial current nonconservation is a pure regularization effect and with another regularization it possibly does not exist. It can be shown however that the condition of the vector current conservation unambiguously determines the amplitude corresponding to the triangle diagram and the expression for the anomaly does not depend on the choice of regularization. In particular anomaly (0.61) can be derived from dispersion relations

when the amplitude is found from its imaginary part. The latter is determined by the unitarity condition and does not need any regularization. In this approach the anomaly is a result of infrared singularity in the theory (Dolgov, Zakharov, 1971).

It is easy to see that the variation of action under transformation (0.59) is

$$\delta S = -\alpha \int \partial_\mu J^A_\mu \, d^4x \tag{0.62}$$

Thus due to the axial anomaly transformation (0.59) gives rise to the terms of the form (0.57) in the Lagrangian. As a result CP-nonconserving terms in the quark mass matrix can be transformed into θ-term or vice versa. So if there existed a massless quark the effects of CP-violation due to L_θ would be unobservable. Experiment however rejects the existence of massless quarks and one has to look for another explanation of the small value of θ like e.g. an existence of a very light pseudogoldstone boson called axion.

G. A short conclusion.

This is in essence the Minimal Standard Model (MSM) of particle interactions and classification. It works perfectly. All the testable predictions of this model are in a very good agreement with experiment. There are however several theoretical problems with the model which remains unclear. First, the forces are not completely unified. There are three independent gauge coupling constants. The ratio of charges of quarks and leptons is not understood. There is no explanation now of very much different masses of the fundamental fermions which span the range from $m_e = 0.5$ MeV to $m_t \approx 150$ GeV or even from m_ν which is at most a few eV for ν_e (and also for ν_μ and ν_τ if they are stable).

What is absolutely mysterious is that why the Nature is described by renormalizable theories. It might be that at the Planck scale all possible kinds of the interactions are permitted but at small energies only the renormalizable ones survive because they fall off with energy only as log of the latter while nonrenormalizable ones go down as a power of energy. Unfortunately this program was not consistently realized. An interesting feature of the standard model is that the quarks and leptons conspire so that there is no anomaly in the physical currents that is in currents which interact with the physical gauge bosons. If this were not true renormalizability of the theory would be broken though in a rather high order in perturbation theory. On the other hand the existence of anomaly in currents which are not coupled to

gauge bosons may be very essential for our existence because it may be the source of the baryon asymmetry of the Universe (see below).

To conclude we do not see any experimental data obtained in a laboratory (in contrast to astronomy or cosmology) which would force us to extend the MSM but there are a lot of problems which are poorly understood theoretically and make the existence of some new physics very much desirable. Cosmology also acts in the same direction but by very much different reason. Theorists-cosmologists would be happy to have as conservative theory as possible and the standard cosmological model probably does not have any serious theoretical drawbacks but experiment or better to say observations are indicating on something new outside the normal physics.

Chapter 3

The Standard Model of the Universe Evolution

The standard big bang theory is in essence the extrapolation into the past of the observed picture of the expanding homogeneous and isotropic (on large scales) Universe. The cornerstones of the theory are the equations of General Relativity (GR) and the equation of state of the gravitating matter. The model is characterized by a small number of measured in astronomy parameters and its consequences, though not numerous, well agree with the observations. The most important evidence in favor of the hot universe model is the microwave background radiation (Penzias and Wilson. 1965) and the predictions of the primordial nucleosynthesis theory (Wagoner, Fowler, and Hoyle, 1967). This makes the standard model unbreakable at least in the space-time region where the corresponding processes take place. Modifications of the standard scenario are possible so far as they do not change the successful predictions of the theory. During the last 15 years the notion of the standard scenario has essentially changed but all this refer to much earlier period than the primordial nucleosynthesis. Now the standard scenario includes the baryosynthesis (Sakharov, 1967), inflation (the history of the problem is given below but probably the first clear formulation of the model was proposed by Guth, 1981), phase transitions and the change of the symmetry of the state with decreasing temperature (Kirzhnits, 1972), and so on. We have mentioned here only the pioneering papers, more references are given in the corresponding sections below.

In principle the large scale distribution of matter and the global geometry

of the space-time may differ from those of the simple standard picture. In particular some inflationary models predict very inhomogeneous universe however on the scales much larger then the present-day horizon, $l > 10^{10^6}$ years (Linde, 1986; Starobinsky, 1986). Moreover the island universe model (Dolgov and Kardashev, 1986) predicts very inhomogeneous distribution of the baryonic matter on the scale of the present-day horizon still compatible with observations. Nevertheless the simple homogeneous and isotropic Friedman cosmology has its region of application.

For the reader interested in the history of big bang cosmology a very good review by Alpher and Herman (1988) may be recommended. The authors of the review together with Gamow collaborated in the formulation of the hot universe model.

3.1. The expansion law in the homogeneous and isotropic model.

The observed homogeneity and isotropy of the Universe on the scales of several hundreds megaparsec permits to take as a good idealization the homogeneous and isotropic model. In such a model the 3-dimensional space is the space of constant curvature. The element of length squared in this space has the form

$$dl^2 = \frac{dx_1^2 + dx_2^2 + dx_3^2}{\left[1 + k\frac{x_1^2+x_2^2+x_3^2}{4a^2}\right]^2} \tag{0.1}$$

where a has the dimension of length and do not depend on the space coordinates x_i but can depend on time. Parameter k can have values ± 1 or 0. If $k = 0$ the space is Euclidean and the cases $k = +1$ and $k = -1$ correspond to the space with constant positive or negative curvature respectively. For $k \neq 0$ a is the curvature radius of the space. If $k = 0$ the scale a is arbitrary. In what follows it is convenient to pass to dimensionless comoving coordinates $\tilde{x}_i = x_i/a$.

In the coordinate frame moving with the matter (the comoving frame) the 4-dimensional interval can be written as

$$ds^2 = dt^2 - a^2 \frac{d\tilde{x}_1^2 + d\tilde{x}_2^2 + d\tilde{x}_3^2}{1 + \frac{k}{4}(\tilde{x}_1^2 + \tilde{x}_2^2 + \tilde{x}_3^2)} \tag{0.2}$$

The metric corresponding to this interval is called the Friedman-Robertson-Walker metric.

For the homogeneous and isotropic distribution of matter the energy-momentum tensor has the form

$$T_{00} = \rho \tag{0.3}$$

$$T_i^j = -p\delta_i^j, (i,j = 1,2,3) \tag{0.4}$$

Here ρ is the energy density and p is the pressure density. In this case the Einstein equations are reduced to the following two equations (see e.g. Zeldovich and Novikov, 1975 or Weinberg, 1972):

$$\ddot{a} = -(4\pi G/3)(\rho + 3p)a \tag{0.5}$$

$$\frac{\dot{a}^2}{2} - \frac{4\pi}{3}G\rho a^2 = -\frac{k}{2} \tag{0.6}$$

where G is the gravitational coupling constant, $G = 6.673 \cdot 10^{-8} cm^3 g^{-1} s^{-2} = 0.59 \cdot 10^{-38} m_p^{-2}$, and m_p is the proton mass. Hence the Planck mass defined as $G = m_{Pl}^{-2}$ is

$$m_{Pl} \equiv G^{-1/2} = 1.30 \cdot 10^{19} m_p = 1.22 \cdot 10^{19} GeV \tag{0.7}$$

Equations (0.5) and (0.6) are easy to understand in terms of the Newton gravitational theory. Let us consider a sphere with radius $a(t)$ and a nonrelativistic test particle on the border of this sphere. The matter outside the sphere is known not to influence this particle. Equation (0.5) is the second Newton law for the test particle with the only difference that in General Relativity not only mass (energy) gravitates but also pressure does. So instead of the mass of the ball bounded by the sphere, $M = (4\pi/3)\rho a^3$ the quantity $(4\pi/3)(\rho + 3p)a^3$ enters the equation. Equation (0.6) is the energy conservation law of the test particle, the constant $(-k/2)$ being the total energy of the nonrelativistic test particle in units of its mass.

Differentiating eq. (0.6) in time and comparing it with eq. (0.5) we get

$$\dot{\rho} = -3H(\rho + p) \tag{0.8}$$

where $H = \dot{a}/a$ is the Hubble parameter. It is easy to check up that the covariant conservation law $D_\mu T_\nu^\mu \equiv T^\mu_{\nu;\mu} = 0$ in the case of homogeneous and isotropic universe

gives just this equation determining the variation of the energy density with time. We may reverse our arguments. Using the well known law of the energy variation by the pressure force

$$dE = -pdV \tag{0.9}$$

and the equation $dV/dt = 3HV$ we obtain the general relativistic conservation law (0.8) and from it and the Newtonian energy balance (0.6) we derive the general relativistic equation of motion (0.5). What is surprising in this argument, that we never used general relativity but get ultimately the correct gravitational equation with gravitating both energy and pressure.

A very important cosmological quantity is the so called critical or closure energy density

$$\rho_c = 3H^2/8\pi G \equiv 3H^2 m_{Pl}^2/8\pi, \tag{0.10}$$

Introducing it into eq.(0.6) we can rewrite the latter as

$$\rho = \rho_c + 3k/8\pi G a^2 \tag{0.11}$$

The present-day value of the critical density is

$$\rho_c = 3H_0^2/8\pi G = 1.87 \cdot 10^{-29} h_{100}^2 \frac{g}{cm^3} = 10.6 h_{100}^2 \frac{KeV}{cm^3} = 1.06 \cdot 10^{-46} h_{100}^2 m_p^4 \tag{0.12}$$

where h_{100} is the dimensionless value of the present day Hubble parameter H_0 measured in 100km/sec/Mpc, $h_{100} = H_0/100 km/sec/Mpc$, and m_p is the proton mass.

Equation (0.11) determines the evolution of the important cosmological parameter $\Omega = \rho/\rho_c$:

$$\rho a^2(\Omega - 1) = \frac{3k}{8\pi} m_{Pl}^2 = const \tag{0.13}$$

Note that if ρ decreases with increasing a faster than a^{-2} then in the course of the Universe expansion Ω deviates further and further away from unity and on the

opposite if ρ decreases slower then a^{-2} then in the course of expansion Ω approaches unity.

Thus there are two equations for three independent functions $a(t)$, $\rho(t)$, and $p(t)$. The missing equation is usually the equation of state $p = p(\rho)$. Different equations of state determines different regimes of the universe evolution. The equation of state is not always applicable however and generally speaking pressure density is not always a function of energy density. For example such a situation is realized in the very simple case of homogeneous scalar field $\phi(t)$.

For gas of nonrelativistic particles the equation of state with a very good precision is the following

$$p = 0 \tag{0.14}$$

(to be more exact the relation between the pressure and energy densities is $p \sim (T/m)\rho \ll \rho$). For $p = 0$ it follows from eq. (0.8) that $\rho \sim a^{-3}$. This result is evident since the energy density of nonrelativistic particles is proportional to their number density.

The expansion law of nonrelativistic matter is especially simple if $\Omega = 1$:

$$a(t) = a_0 \cdot (t/t_0)^{2/3} \tag{0.15}$$

For the known function $\rho(a)$ equation (0.6) can be integrated for arbitrary Ω. The duration of the expansion t_U (the Universe age) is expressed through the corresponding to that moment values of the Hubble parameter H and Ω as follows

$$t_U = f(\Omega)/H \tag{0.16}$$

where

$$f(\Omega) = \Omega(\Omega - 1)^{-3/2}\left[\arcsin\frac{\sqrt{\Omega - 1}}{\Omega} - \frac{\sqrt{\Omega - 1}}{\Omega}\right] \tag{0.17}$$

It can be easily seen that $f(1) = 2/3$ in accordance with eq (0.15) and $f(\Omega \ll 1) \approx 1 + (\Omega \ln \Omega)/2$. As we see in what follows the dependence of the Universe age t_U

on H and Ω is used for the derivation of the bounds on the relic particle masses (Gerstein and Zeldovich, 1966). To this end the following approximate expression for t_U is convenient

$$t_U \approx H^{-1}\left(1+\frac{\sqrt{\Omega}}{2}\right)^{-1} \approx 0.98\cdot 10^{10} h_{100}^{-1}\left(1+\frac{\sqrt{\Omega}}{2}\right)^{-1} \; years \tag{0.18}$$

It is valid in the range $0 \leq \Omega \leq 4$ with the accuracy better then 0.05.

For relativistic ideal gas the equation of state is known to have the form (see e.g. Landau and Lifshits, 1964):

$$p = \rho/3 \tag{0.19}$$

The energy density in this case scales as a^{-4} (see eq. (0.8)). The extra power of a^{-1} in comparison with the nonrelativistic case is due to the fact that not only the particle number density decreases in the course of expansion but also the average energy of the particles goes down as a^{-1}. Now the expansion time is connected with the running values of Ω and H by the relation

$$t = h^{-1}(1+\sqrt{\Omega})^{-1} \tag{0.20}$$

If $\Omega = 1$ the expansion is also of the power law form, as in the nonrelativistic case (0.15), but with the different power

$$a(t) = a_0(t/t_0)^{1/2} \tag{0.21}$$

If $\Omega \leq 1$ and the Universe is open the expansion will proceed infinitely long while for $\Omega > 1$ (closed Universe) the expansion shall turn into contraction. However this connection between the Universe curvature and her evolution is true only if the energy density decreases faster than a^{-2} (see eq. (0.6)).

Since in the course of expansion the energy density of relativistic matter decreases faster then that of relativistic matter the early Universe should be dominated by relativistic matter. This period is called RD-stage while the period of nonrelativistic matter dominance is called MD-stage. This period is probably realized now

but the dominance of relativistic matter in the present stage is not excluded. To determine the moment of the transition from RD-stage to MD-stage let consider the following two simple analytically solvable models with $\Omega = 1$.

Let the Universe be filled by gas of noninteracting particles with mass m. The number density of these particles at moment t_1 is equal to N_1 and their energy is $E_1 = (p_1^2 + m^2)^{1/2}$. The energy density at the moment corresponding to the scale factor a is

$$\rho = N_1 \left(\frac{a_1}{a}\right)^3 \left(m^2 + \frac{p_1^2 a_1^2}{a^2}\right)^{1/2} \tag{0.22}$$

The account is taken of the change of the number density because of the rise of the volume and the red shifting of the momentum due to expansion. The latter will be more clear after acquaintance with the next section. Substituting this expression into eq. (0.6) we find the expansion law:

$$t = \tau_\nu \left[\left(\frac{m^2 a^2}{p_1^2 a_1^2} + 1\right)^{3/4} - 1\right] \tag{0.23}$$

where $\tau_\nu = (m_{Pl}/m^2)(p_1^3/6\pi N_1)^{1/2}$ is the characteristic time scale of the transition from one stage to another. Note that it does not depend on the moment t_1 when the initial conditions are fixed because the ratio (p^3/N) does not change in the course of expansion.

The Hubble parameter in this model is

$$H \equiv \frac{\dot{a}}{a} = \frac{2}{3} \frac{(t + \tau_\nu)^{1/3}}{(t + \tau_\nu)^{4/3} - \tau_\nu^{4/3}} \tag{0.24}$$

This model approximately describes the Universe expansion when the energy density is dominated by massive neutrinos. Of course the energy of a neutrino is not a constant quantity but has the Fermi distribution with a decreasing during expansion temperature (see below). Nevertheless the results obtained are approximately valid if E_1 is the average neutrino energy $\langle E_1 \rangle \approx 2.7E$ and N_1 is their number density, $N_1 \approx 0.09T^3$ (all for $T \gg m$). This results in the following value of τ_ν

$$\tau_\nu \approx 3.4 m_{Pl}/m_\nu^2 = 3 \cdot 10^{11} sec(10eV/m_\nu)^2 \tag{0.25}$$

At $t = \tau_\nu$ the scale factor differs from its contemporary value by

$$z_\tau + 1 \equiv a(t_U)/a(\tau_\nu) \approx (t_U/\tau_\nu)^{2/3} \approx 10^4 (m_\nu/10eV)^{4/3} \tag{0.26}$$

This justifies the used above assumption that $\Omega = 1$ at $t = \tau_\nu$.

Let us consider now another model when the matter consists of two components: massless particles (photons) with energy density ρ_r and nonrelativistic ones (baryons) with energy density ρ_m. It is assumed that initially at moment $t = t_1$ $\rho_m/\rho_r \equiv \varepsilon \ll 1$ $t = t_1$ and as before that $\Omega = 1$. Equation (0.6) also can be integrated and the expansion law is of the form:

$$t = \tau_B \left[\frac{2}{3} + \frac{1}{3} \left(1 + \frac{a}{l}\right)^{3/2} - \left(1 + \frac{a}{l}\right)^{1/2} \right] \tag{0.27}$$

where $l = a_1/\varepsilon$, $a_1 = 2t_1$ is the value of the scale factor at the moment t_1, and $\tau_B = 2l/\varepsilon = 4t_1/\varepsilon^2$. τ_B can be expressed through the ratio of the number densities of photons and baryons:

$$\tau_B = 10^{13} sec(m_B/1GeV)^2 (N_\gamma/10^9 N_B)^2 \tag{0.28}$$

The result evidently does not depend on the choice of t_1.

The Hubble parameter is equal to

$$H = 2\tau_B^{-1} \left(1 + \frac{a}{l}\right)^{1/2} \left(\frac{l}{a}\right)^2 \tag{0.29}$$

Equations (0.27) and (0.28) parametrically determine the function $H(t)$.

In both cases that we have just considered the change of the regime proceeds rather slowly, the two-component matter becomes nonrelativistic even slower then the gas of massive neutrinos.

The considered up to now equations of state $p = \rho/3$ and $p = 0$ do not exhaust all physically interesting cases of the linear dependence

$$p = \gamma\rho \tag{0.30}$$

This equation leads for $\Omega = 1$ to the following expansion law

$$a \sim t^{2/3(\gamma+1)}, \tag{0.31}$$

$$H = 2/3t(\gamma + 1), \tag{0.32}$$

$$\rho = [6\pi G(\gamma + 1)^2 t^2]^{-1} \tag{0.33}$$

Physical objects which satisfy eq. (0.30) for $\gamma = 1, 1/3, 0, (-1/3), (-2/3)$, and (-1) are known. (Why $\gamma = 2/3$ is absent in this sequence?)

The case $\gamma = -1$ corresponds to the stiffest equation of state (Zel'dovich, 1961), $\gamma < -1$ is impossible because this gives the speed of sound larger then the speed of light. The relation $p \approx \rho$ can be realized for a scalar field during cosmological contraction The cases $\gamma = -1/3$ and $\gamma = -2/3$ correspond respectively to the dominance of the chaotic gas of cosmic string and domain walls.

Very interesting for what follows is the point $\gamma = 1$:

$$p = -\rho \tag{0.34}$$

Equations presented above are not applicable in this case. Instead of them we get

$$H = const, \tag{0.35}$$

$$a = a_0 \exp(Ht) \tag{0.36}$$

$$\rho = 3H^2/8\pi G \tag{0.37}$$

It is assumed that $\Omega = 1$.

Such regime of expansion arises when the source of gravitational field has the form

$$T_{\mu\nu} = \rho_{vac} g_{\mu\nu} \tag{0.38}$$

This quantity in fact coincides with Λ-term introduced by Einstein into gravity equations in 1918. It is natural to call ρ_{vac} the energy density of vacuum. Astronomical observations give the following bound on the value of the vacuum energy

$$\mid \rho_{vac} \mid < \rho_c \approx 10^{-46} m_N^4 \tag{0.39}$$

We will see in what follows that a scalar field on the stage of expansion could lead to the relation $p = -\rho$ which effectively corresponds to the vacuum energy.

The metric corresponding to the gravitating vacuum (or to nonzero cosmological constant) is called the De Sitter metric. Equations (0.35,0.36,0.37) describe spatially flat De Sitter world. The solutions of the gravity equations for open and closed De Sitter space are respectively

$$a(t) = H_v^{-1} sinh(H_v t), H(t) = H_v cth(H_v t) \tag{0.40}$$

$$a(t) = H_v^{-1} cosh(H_v t), H(t) = H_v th(H_v t) \tag{0.41}$$

where $H_v = (8\pi G \rho_{vac})^{1/2}$.

In contrast to the Friedman cosmology even the closed world with $\rho_{vac} > 0$ expands eternally. This is due to antigravity generated by the space components of $T_{\mu\nu}$. This statement may need more clarification. We used to say that gravitational interaction can be only attractive because it is realized by the spin-two particles and the energy is positive definite. On the other hand the source of gravitational acceleration, as one sees from eq.(0.5) is not energy density but $(\rho + 3p)$. This quantity may have either sign. Still for the objects which are localized in space, gravitational interactions are always attractive. To see it let us consider a weak gravitational field and integrate in 3 dimensional space the quantity $x^j T^\mu_{\nu;\mu}$. This is zero since T^μ_ν is conserved. We may integrate this zero by parts and get

$$\int d^3x x^j T^\mu_{\nu;\mu} = -\int d^3x T^j_\nu = 0 \tag{0.42}$$

Thus one sees that for local object, for which this integral converges at infinity, the space components of the energy-momentum tensor do not participate in creating

gravitational field and it is only energy density which gravitates. However if the integration by parts is impossible gravity may be repulsive. In particular in the case of nonzero vacuum energy $\rho + 3p = -2\rho < 0$ and the expansion proceeds with positive acceleration in contrast to the normal equation of state when the acceleration is negative. This may be the source of the initial push which resulted in the Universe expansion which we see today.

The Universe age for the case when all the mentioned above forms of the gravitating matter are essential (except for those giving the stiffest equation of state), can be found with the help of eq. (0.6):

$$t_U = H^{-1} \int_{x_{min}}^{1} dx x[\Omega_r + \Omega_m x + \Omega_s x^2 + \Omega_w x^3 + \Omega_v x^4 + (1 - \Omega_{tot})x^2]^{-1/2} \quad (0.43)$$

where Ω_a is the relative part of the energy density (in terms of ρ_c) which is taken by relativistic gas (r), nonrelativistic gas (m), strings (s), domain walls (w), and vacuum (v). The lower integration limit is determined by positiveness of the integrand. Evidently in the corresponding limiting cases we get equations (0.16) and (0.20).

The measured values of the Hubble constant, $H = 50 - 100$ km/sec/Mpc are compatible with the data on the Universe age, $t_U > 12 \times 10^9$ years for H at the lower end of the permitted region. The problems worsens with large Ω, which is predicted by inflationary models to be precisely 1 (with accuracy about 10^{-4}). So we must either conclude that the Hubble constant is rather small or that gravity of vacuum or of other exotic objects is essential in the Universe.

3.2. Problems of the Friedman Cosmology and the Idea of Inflation.

Homogeneous and isotropic cosmological model astonishingly well describes the observed properties of the Universe. However this model could be realized only with a perfect fine-tuning of the initial conditions. With the slow power law expansion only negligibly small manifold of possible initial states could result in the observed today Universe. One may think that the initial conditions were realized accidentally and the question about their origin is not sensible. A decade ago just this point of view dominated. Now the situation has completely changed. It became clear that the initial conditions for the Friedman cosmology were realized dynamically and the creation of the universe of our type is a result of the exponential expansion naturally arising on very early stage of the Universe evolution.

We believe that now the Universe is dominated by nonrelativistic matter with the equation of state (0.14) and with the energy density $\rho \sim a^{-3}$. On the earlier stages relativistic gas (0.19) with $\rho \sim a^{-4}$ dominated. The change of the regime took place at the red-shift $z \approx 10^3 \div 10^4$ (see eqs. (0.26) or (0.28)). In accordance with observations the contemporary value of parameter Ω does not much deviate from unity, $\mid \Omega - 1 \mid = O(1)$. Correspondingly due to eq. (0.13) the deviation of Ω from unity *at the earlier stages should be negligibly small*. In particular to the moment $t = 1$sec when the temperature of the primeval plasma was $T = 1MeV = 10^9$K and $z = 3 \cdot 10^9$, the deviation was

$$\mid \Omega(1sec) - 1 \mid = 10^{-15} \div 10^{-16} \tag{0.44}$$

and at the Planck moment $t_{Pl} = 10^{-43}$sec it was

$$\mid \Omega(t_{Pl}) - 1 \mid = 10^{-58} \div 10^{-59} \tag{0.45}$$

This means that in order to survive from the Planck time to the present age $t \approx 10^{18}$sec *the initial state should be prepared* with the fantastic accuracy of about 10^{-60}. If initially $\mid \Omega - 1 \mid = O(1)$ the characteristic Universe life-time would be 10^{-43} sec but not 10^{18} sec.

The same difficulty can be illustrated in a different way. The most distant visible astronomical objects are situated at the distance of the order of $l_U = 10^{10}$years$\approx 3 \cdot 10^{17}$sec. Going back into the past down to $t = t_{Pl}$ we see that the Universe size at that time was about

$$l(t_{Pl}) = \frac{3K}{m_{Pl}} l_U \approx 10^{-14} s \approx 10^{29} l_{Pl} \tag{0.46}$$

It would be natural to think that the Universe size at the Planck time was also of the Planck scale. The difference in 29 orders is "slightly" discouraging.

The possibility of reasonable extrapolation to $t = t_{Pl}$ might be questioned but the extrapolation to $t = 1$sec seems to be perfectly well justified. The weighty argument in its favor is the agreement with observations of the theory of the primordial nucleosynthesis which proceeded just at that time. It is very difficult if at all possible to believe that the accuracy of the fine-tuning of the initial conditions in the Universe was as good as 10^{-15} or even better. Thus one have to conclude

that the expansion law in the past differed from the power law (0.15) or (0.21) and was most probably the exponential one when starting from more or less arbitrary Ω we would come to Ω close to unity with exponential accuracy as it is seen from eq. (0.13). In that sense the exponential expansion in the past is the "experimentally" established fact. The problem of the extreme fine-tuning of Ω at the beginning of the Friedman stage is called

1)the flatness problem. As we have seen it is solved if somewhere in the past the energy-momentum tensor was dominated by the term of the form (0.38) that is by effective cosmological term. Of course the real cosmological constant does not change with time because the covariant $T_{\mu\nu}$-conservation leads to the condition $\rho_{vac} = const$ and the contemporary value of ρ_{vac} is too small (if nonzero) for the solution of the problems we are interested in. Thus the state of matter which produce the energy-momentum tensor approximately coinciding with (0.38) should be unstable and ultimately evolve to the normal state. Such a regime can be realized e.g. by a scalar field with flat potential or by the real initial Λ-term somehow compensated during the later Universe evolution. The adjustment mechanism naturally gives rise to the time-dependent cosmological term (Dolgov, 1982). Recall in this connection that vacuum energy in the past could differ from the contemporary value because of the phase transitions in the primeval plasma when the temperature went down (Kirzhnits, 1972; Kirzhnits and Linde, 1972). Immediately the question arises why the vacuum energy is compensated with such a fantastic precision. Thus we come to the

2).problem of the cosmological constant. In accordance with theoretical estimates the change of vacuum energy during the possible phase transitions in the early Universe is much larger then the upper limit for its contemporary value (0.39). For example the phase transition in QCD from free quarks and gluons at high temperatures, $T > 100$MeV, to the confinement phase is accompanied by formation of e.g. gluonic condensate (Vainshtein, Zakharov, and Shifman, 1978) with vacuum energy

$$\rho_{QCD} \approx 10^{-4}(GeV)^4 \tag{0.47}$$

This is 50 orders of magnitude larger then ρ_{vac} at the present time.

The electroweak phase transition from the $SU(2) \times U(1)$-symmetric phase to the phase with broken symmetry, when the condensate of the Higgs field is formed, leads to the change in vacuum energy equal to

$$\Delta\rho_{EW} \approx 10^{8}(GeV)^{4} \tag{0.48}$$

If the grand unification models are valid than the analogous phase transition gives

$$\Delta\rho_{GUT} \approx 10^{60}(GeV)^{4} \tag{0.49}$$

Apart from that the quantum zero mode fluctuations contribute into ρ_{vac}. The contribution of a separate field is infinite and it was usually simply subtracted out because in the absence of gravity the origin of the energy scale is unessential. The account of gravity changes the situation and zero mode fluctuations create a problem. Supersymmetric theories slightly improve the case because in the limit of exact supersymmetry the contributions of fermionic and bosonic zero modes are canceled out. This was noted by Zel'dovich (1968) before the advent of supersymmetric theories. We know however that supersymmetry is not exact and hence the zero mode oscillations do not canceled out exactly but only down to the terms of the order of m_{SUSY}^{4} where m_{SUSY} is the scale of the supersymmetry breaking. If a global supersymmetry is spontaneously broken the vacuum energy must be nonzero. In supergravity a nonvanishing ρ_{vac} is not obligatory and one can adjust the parameters so that it vanishes when the supergravity is broken but the required fine-tuning is unnatural and should be accurate with the precision of about 10^{-100}.

So we see that the changes in the vacuum energy during the Universe evolution could be 50-100 orders of magnitude larger then the bound on the value of ρ_{vac} at the present time. It is hard to imagine that the initial value of ρ_{vac} was adjusted in such a way that the subsequent phase transitions canceled it out with this fantastic accuracy. There must exist a mechanism of the vacuum energy compensation. We discuss the possible ways of the solution of this central cosmological problem below and now describe some other, maybe not so impressive, but still essential cosmological problems.

3) Isotropy and homogeneity of the Universe in the Friedman cosmology were understood as result of the choice of some particular initial conditions the origin of which was until recently quite mysterious.

4) The horizon problem is close to the problems of isotropy and homogeneity and is created by the fact that the background electromagnetic radiation coming from different parts of the sky looks absolutely the same. Without the dipole asym-

metry connected with the Earth movement with respect to the background the relative angular fluctuations of the radiation temperature is smaller then 10^{-4}. On the other hand the regions in the sky with the angular size smaller then 0.03 is casually disconnected. Indeed the background radiation stopped to interact with the matter after the hydrogen recombination at $T \approx 3000$K i.e. at $z_r = T_r/T_0 \approx 10^3$. This corresponds to $t_r \approx 10^{13}$sec from the beginning. Hence the size of the region where photons could interact in the modern scale is smaller than $R_r = z_r t_r \approx 10^{16}$sec. This is approximately 30 times smaller than the visible size of the Universe. These arguments show that there should exist a mechanism which had burned all the Universe simultaneously in the regions which for the usual expansion mechanism could not exchange the light signals. The discussed below exit from the inflationary stage gives, as we see, the example of such mechanism.

5) Large scale structure of the Universe means to the galaxies, their clusters, and superclusters on the homogeneous background in the very large scale. It is assumed that the observed picture was created due to gravitational instability of initial fluctuations of density. A natural candidate for such inhomogeneities are quantum field fluctuations. Their amplitudes and the characteristic scales are however too small to create the structure in the usual regime of expansion. It can be shown that inflationary regime gives rise not only to the fluctuations with sufficiently large wave length (this is evident) but also to large amplitude of the fluctuations. In a sense the plan is over fulfilled since in natural models the amplitudes happen to be too large. Alternative sources of the initial density fluctuations could be domain walls, cosmic strings, and other topological or nontopological solitons.

In accordance with the theory of gravitational instability the latter is developed when in the course of expansion nonrelativistic matter becomes dominating and pressure or free streaming out does not prevent from gravitational clumping. The onset of instability in usual baryonic matter takes place rather late after hydrogen recombination when the radiation pressure becomes sufficiently small. The observed anisotropy of the background radiation in this case puts a very stringent upper limit on the amplitude of the energy density fluctuations and the time left for the development of the perturbations up to the observed today magnitude proves to be too short. Theory of the large scale structure formation in the Universe could overcome this difficulty if along with baryons there existed some other particles noninteracting with electromagnetic radiation e.g. massive neutrinos. If this is the case the neutrino mass should be a few tens of electronvolts. However the neutrinos alone cannot quantitatively describe the observed large scale structure. Fortunately there are plenty of other candidates for the role of creators of the Universe structure

e.g. superpartners like gravitino, photino, etc., axion, or some more exotic objects predicted by the modern proliferation of the particle physics models. One of the popular today hypothesises is that 70% of the Universe mass is contained in new heavy (supersymmetric) particles and 30% in light ($m \approx 10$ eV) neutrinos. Competing proposals are neutrinos plus vacuum energy, or neutrinos plus astronomically heavy seeds, and so on.

The new collisionless particles could solve

6)the problem of the hidden mass of the Universe. Observations show that there is a lot of matter in the Universe (about 90%) which is not seen. This is the so called dark matter which makes the hidden mass. It seems practically certain that the hidden mass cannot be attributed to the usual baryonic matter and besides weakly interacting particles there is only exotic possibility of modification of gravitational interaction at large distance.

7)The problem of the baryon asymmetry of the Universe that is the problem of existence of matter and the absence of antimatter in the observed world seems to be settled down now. If particle interactions are charge asymmetric (we know from experiment that this is indeed the case) and baryonic charge is not conserved then in nonstationary situation there should be generated an excess of particles over antiparticles or vice versa. The sign usually cannot be predicted, but the amplitude is evaluated to be of the proper order of magnitude.

It is interesting to note in this connection how our attitude to the proton stability has changed. Earlier it was implicitly assumed that our existence proves that proton is stable but now the conclusion is just the opposite: our existence proves that proton is unstable or to be more exact that baryonic charge is not conserved. Indeed otherwise the baryonic asymmetry would not be generated and the Universe would be quite different and not suitable for life.

In theories of grand unification together with baryonic charge nonconservation the existence of magnetic monopoles is predicted. The number density of the latter in the Universe is calculated to be unacceptably large if the expansion law is of the standard Friedman form (Zel'dovich and Khlopov, 1978; Preskill, 1979). The discrepancy between the theory and the observational bounds is about 10 orders of magnitude. This contradiction creates

8)the magnetic monopole problem.

Possibly the list should include also

9)the problem of the space-time dimension. Why $D = 4$, though any other D seem possible? Standing on the anthropic principle it is possible to justify $D = 4$ because life, in our understanding of it, is possible neither in $D > 4$ nor in $D < 4$ Still it is more attractive to have the dynamic solution of the problem. Probably specific classical field configuration which could exist only in $D = 4$ permit only three space dimensions to become large.

And at last but not the least there is the most fundamental problem not only in cosmology but in physics in general:

10) the problem of creation and ultimate fate of the Universe. In popular lectures this problem is sometimes formulated as "Was there the beginning and will there be the end of the world?" It is difficult to realize that time could be finite, so psychologically the model of eternal Universe seems more attractive. One possible model of this kind is the model of oscillating Universe in which expansion and contraction phases alternate for infinitely long time. Such a Universe should be closed, that is $\Omega > 1$. Unfortunately rather general arguments show that in oscillating universe the entropy should rise due to particle production by gravitational field near singularity. This would result in the infinite increase of the universe radius in the infinite time. Moreover the problem of passing through the singularity remains open.

Rather popular in the recent years was the idea of the Universe creation from nothing (Tryon, 1973; Fomin, 1975). There is no contradiction with the energy conservation law in this process if a closed universe is created. For the latter the total mass is known to be zero because of the gravitational mass defect. The formal description of the universe creation resembles the process of the electron-positron pair production from vacuum by electric field. Still the final theory of the process is not developed.

It is noteworthy that the created from nothing a closed universe with the Planck size could reach the present state in a finite number of oscillations because of the increase of entropy.

Eternally existing universe can be based on the scalar field with infinitely many decreasing local minima in the potential energy not bounded from below (Dolgov, 1985). In this model the universe evolution would look as the infinite

sequence of big bangs with subsequent expansion and cooling down and the next explosion and so on.

An interesting model of eternal universe has been considered by Linde (1986) in the framework of the inflationary scenario when the Universe, as a whole, mostly stays in quantum state and only little pieces of it becomes classical due to fluctuations of a scalar field which inflates them to a classical size. Though this picture creates a lot of questions, like the meaning of the quantum space-time and the law of the evolution of the latter, it may be an interesting alternative to the scenario of the creation of the Universe from nothing.

The mentioned possibilities are sooner illustrative than serious and now we do not have a reliable answer to the question about the Universe creation and probably are far from it. One should remember however that quite recently many of the problems discussed in this section seemed to be outside the scope of science and now the beautiful solution has been found based only on the assumption that at an early stage of the Universe evolution the scale factor exponentially rose with time, $a(t) \sim \exp(H_v t)$.

This simple assumption permits to solve problems 1,3,4,8, and to some extent 5 in a unified way. In this scenario the Universe on the inflationary stage was exponentially expanding emptiness. All initially existing matter density quickly tended to zero and the initial conditions were forgotten with exponential precision. The space became as homogeneous and isotropic as vacuum could be. The parameter Ω tended to the closure value in accordance with the law

$$(\Omega - 1) \sim \exp[(-8\pi\rho_{vac}/3m_{Pl}^{1/2})t] \equiv \exp(-H_v t) \qquad (0.50)$$

The necessary duration of the exponential stage is given by the condition

$$H_v T > 70 - \ln(m_{Pl}/T) \qquad (0.51)$$

where T_v is the temperature of the primeval plasma after the vacuum explosion, reheating, and the transition to the Friedman expansion regime (see below). Condition (0.51) ensures $\Omega = O(1)$ at the present time as it is demanded by the existence of our world. Correspondingly at $T = 1$MeV the condition $\mid \Omega - 1 \mid \leq 10^{-15}$ is satisfied (eq. (0.44)).

Evidently the inflationary model solves the problem of magnetic monopoles if the exponential expansion took place after the phase transition at which the monopoles were formed or if the reheating temperature was smaller then the monopole mass.

The horizon problem is also naturally solved in the model if the visible now part of the Universe was a microscopic casually connected region before the inflation.

For the concrete model of inflationary Universe two problems are essential:

1) Mechanism of inflation or in other words the mechanism of formation and disappearance of the vacuum or vacuum-like energy.

2) The end of inflation i.e. the transformation of the vacuum energy into the energy of matter (of elementary particles) and the onset of the Friedman expansion regime.

No other way to solve the fundamental cosmological problems enumerated in this section is known and it seems certain that the Universe have undergone the exponential expansion stage and thus inflationary cosmology is now a firmly established part of the standard model. Still no "no-go" theorems are known and one cannot exclude future competing proposals. Crucial for the inflationary scenario would be an accurate measurement of Ω which should be very close to 1 (at least in the traditional versions of the model).

3.3 Short history of the Universe.

After our Universe has been somehow created we can more or less definitely describe her subsequent evolution. Most probably initial values of appropriate physical quantities were close to the Planck values:

$$t_{Pl} = m_{Pl}^{-1} \approx 5 \cdot 10^{-44} sec \tag{0.52}$$

$$l_{Pl} = m_{Pl}^{-1} \approx 1.5 \cdot 10^{-33} cm \tag{0.53}$$

$$\rho_{Pl} = m_{Pl}^{4} \approx 2 \cdot 10^{76} GeV^4 = 2.5 \cdot 10^{117} GeV/cm^3 = 4 \cdot 10^{93} g/cm^3 \tag{0.54}$$

The physical mechanism of the initial push which has caused the Universe expansion was unknown till the recent time. Now it is believed that the observed

Hubble flow was generated by the antigravity created at the very early stage of the Universe evolution by the vacuum-like state of matter with the energy momentum tensor $T_{\mu\nu} \sim g_{\mu\nu}$ Such energy-momentum tensor induces exponential expansion and the arguments of subsec. 3.2 strongly suggest that this regime did existed.

The part of the Universe which underwent exponential expansion rose up to fantastically large size generically much larger than the present-day horizon. The only form of energy except for real vacuum energy which could survive in this enormously expanding volume is the energy of homogeneous scalar field which could be the source of inflation. A very important moment in the Universe history is the end of inflation when the vacuum-like energy of the scalar field is transformed into the energy of the plasma of elementary particles. Everything that was before that moment our Universe effectively forgets. The only reminiscences of inflation which remain are isotropy, homogeneity, flatness, and as we see in what follows small perturbations of density and metric. All the relic particles which existed before inflation disappear. Note however that together with grand inflation which solves the flatness problem there could be shorter exponential stages which probably arose due to supercooled phase transitions of the first order at later stages.

After inflation was over the primeval plasma of elementary particles quickly evolves to the thermal equilibrium state. All types of particles with $m < T$ are produced. An exception may be particles with anomalously weak interactions like gravitons or gravitinos which possess only gravitational interaction. Their production can be rather small and they never come into the equilibrium with the cosmic plasma.

At some period after inflation there should proceed baryosynthesis resulting in excess of baryons over antibaryons. It takes place between 10^{16}GeV and 10^{2}GeV depending upon the model.

The value of the baryonic charge density puts a limit on the possible entropy rise due e.g. to supercooling in the phase transitions after the baryosynthesis. On the other hand the phase transition producing magnetic monopoles must take place before inflation was over so that their number density would be acceptably small.

Phase transitions in the primeval plasma is a new feature of the cosmology of the early Universe which came to cosmology with the advent of spontaneously broken gauge theories (see sec.II). The macroobjects formed in these phase transitions like cosmic strings or domain walls can have a strong influence on the Universe evolution.

Near a phase transition the equation of state is rather complicated and the expansion law differs from the simple power one. However generically phase transitions last for a rather short time and the equation of state soon returns to the usual equation of state of relativistic ideal gas, $p = \varepsilon/3$ and $a \sim t^{1/2}$. It is also possible that at some stages the energy density was dominated by long-lived nonrelativistic particles which had got their masses as a result of the phase transition and the equation of state became close to $p \ll \varepsilon$ and so $a \sim t^{2/3}$.

We believe that there existed the phase transition on the grand unification scale at $T = 10^{15}$GeV which has given masses to the gauge X- and Y-bosons (leptoquarks). The order of this phase transition is not known. It depends on the Higgs sector of the theory and could be both of the first and of the second order. In the first inflationary models the theory was specially chosen so that the grand unification phase transition was of the first order with a strong supercooling ensuring a long dominance of the vacuum-like energy.

Going down to the temperatures about 10^2GeV we come on a more firm ground of the particle theory confirmed by experiment. In this temperature range the transition to the nonsymmetric phase of electroweak interactions takes place and W and Z become massive. The order of the electroweak phase transition is not known. Most probably it is the second order (especially if minimal standard model with a heavy Higgs boson is true), but more work is necessary to clarify the situation. This issue is of primary importance for the electroweak baryogenesis.

And last but not the least at $T = 100 \div 200$MeV there is the phase transition from the phase of free quarks to the confinement phase. Because of the large value of the coupling constant at this energy scale the theory of the phase transition proves to be very complicated and its character is unknown. The restriction on the entropy increase demands that it should be of the second or of weakly first order.

At $T < 100$MeV the primeval plasma consists only of photons, e^-, e^+, and of three(?) types of neutrinos and a small admixture of nucleons and possibly some other (massive) relic particles which are yet unknown.

At $T = 3 \div 5$MeV neutrinos stops to interact with the plasma and freely propagate in the Universe. Note however that the high energy neutrinos decouple later and this results in the distortion of the neutrino spectrum.

When the temperature drops below 1Mev the weak $(n - p)$-transitions are

switched off and the ratio n/p freezes. The value of the latter determines the light element abundances which have been produced during primordial nucleosynthesis. Starting from this moment the predictions of the standard cosmological model is quantitatively checked by observations.

Primordial nucleosynthesis proceeds at the temperatures about 100KeV when the light elements such as $^2H, ^3H, ^3He, ^4He, ^7Li$, and so on are produced. The chain of reactions leading to the formation of these nuclei is well determined. The first calculation of light element production has been performed by Wagoner, Fowler, and Hoyle (1967) and the results are in beautiful agreement with observations. This agreement does not permit any considerable modification of the standard scenario and serves as a sensible method to get an information about elementary particles from cosmology. In particular hence the celebrated Schwartsman (1969) limit on the number of neutrino species follows. Recently new more advanced numerical codes for the calculation of the primordial light element abundances were worked out and their production is calculated in detail. These calculations in particular presents a strong evidence in favor of nonbaryonic invisible matter in the Universe.

The next important step is the transition from RD-stage to MD-stage. In accordance with estimates of subsec. 3.1 it takes place at $t = 10^{11} \div 10^{12}$sec and correspondingly $T = (3 \div 1) \cdot 10^4 K$. Slightly later when the temperature drops down to 3000K the hydrogen recombination takes place, $e^- + p \to H^0$, and the plasma becomes transparent to the background radiation. ¿From that time the light pressure does not prevent from gravitational clumping of baryonic matter and the epoch of the formation of protostar and protogalaxies begins. Gravitational instability of invisible matter (e.g. of massive neutrinos) should manifest itself earlier with the onset of MD-stage and gives rise to the formation of first gravitationally bound structures. Without these seeds, which are necessary for the subsequent capture of the usual matter, the theory of large scale structure formation in the Universe encounters considerable difficulties. After protostructures were formed they attracted the usual matter and ultimately the beauty of stars and galaxies, which we see now, was shining on the sky.

Chapter 4

Baryogenesis

Astronomical observations show that at least in our neighborhood antimatter in the form of antiprotons, antineutrons, and positrons is practically absent and that most probably all the visible part of the Universe consists of matter. Earlier the excess of matter was considered as one of initial conditions of Friedman cosmology and the value of the ratio of baryonic charge density to number density of photons

$$\beta_0 = (N_B/N_\gamma) \approx 3 \cdot 10^{-10} \tag{0.1}$$

was believed to be one of fundamental cosmological constants imposed "externally".

Now the situation has completely changed. The mechanisms that can produce particle or antiparticle excess from initially charge symmetric state are found. The value of this excess can be expressed through characteristic parameters of microphysics. The history of the problem goes up to the paper by Sakharov (1967) where the basic ideas of baryosynthesis in the Universe have been formulated. After the paper by Kuz'min (1970) who considered slightly different model the subject was forgotten for several years because at that time no theoretical reasons for baryon nonconservation was known. The advent of grand unification models created the necessary basis and later the works by Ignatiev, Krasnikov, Kuz'min, and Tavkhelidze (1978) and by Yoshimura (1978) gave rise to the stream of papers on baryogenesis which has not dried up until now. Moreover recent years showed a burst of activity on the subject connected with the possibility of baryogenesis on electroweak scale. For acquaintance with early papers on the subject the reviews by

Dolgov and Zel'dovich (1981) and by Kolb and Turner (1983) can be helpful. Up to date description of the issue can be found in the review papers by Dolgov (1992) and by Cohen, Kaplan, and Nelson (1993).

Three principles of baryogenesis formulated by Sakharov are the following. First, it is evident necessity of nonconservation of baryons. Otherwise the difference between the numbers of baryons and antibaryons that is baryonic charge cannot change. Second, a violation of symmetry between particles and antiparticles seems also necessary. The transformation of particles into antiparticles is called charge or C-conjugation and strong, electromagnetic and gravitational interactions are invariant with respect to it. Weak interactions permit to distinguish between particles and antiparticles since they break C-parity with relative probability of the order of one. However breaking of C-parity is not enough for generation of excess of particles over antiparticles in the Universe. If the interactions are invariant with respect to the so-called combined inversion CP when charge conjugation is accompanied by the space reflection then no excess of particles over antiparticles can be generated. This statement is practically evident because in this case the probability of any process with particles is equal to the probability of the mirror reflected process with antiparticles and so after averaging over space and particle spins total probabilities of processes with particles and antiparticles coincide. However it is observed experimentally that CP-invariance is broken. Because of C- and CP-violation the probability of charged conjugated processes $i \to f$ and $\bar{i} \to \bar{f}$ are different, $\Gamma_{if} \neq \Gamma_{\bar{i}\bar{f}}$. Though the mechanism of the breaking is not known, there are plenty possible theoretical models which may explain the phenomenon. Our world probably exists only thanks to this small violation of CP-symmetry.

The observed in experiment violation of C, P, and CP well illustrate the validity of the principle "everything that is not forbidden is permitted" in physics. Theory prescribes only invariance with respect to the combined action of three operations C, P, and T where T is the operation of time reversal which interchange initial and final states and change the signs of velocities and angular momenta of particles. CPT-invariance is a consequence of general principles of the theory such as Lorentz-invariance, analyticity, and positive definiteness of energy (see review by Grawert, Luders, and Rollnik, 1959). CPT theorem ensures some symmetry between particles and antiparticles in particular the equality of particle-antiparticle masses and for unstable particles the equality of their total life-times.

The third necessary condition for generation of charge asymmetry is violation of thermodynamic equilibrium (Okun and Zel'dovich, 1976). As is well known

(see e.g. Landau and Lifshits, 1964) equilibrium distribution functions are determined only by energies and chemical potentials of particles. If the particles do not posses any conserved charge their chemical potential in equilibrium vanishes and the numbers of particles and antiparticles are equal. Note the role of CPT-theorem in this statement: it demands the equality of particle-antiparticle masses. Thus starting from a state with nonzero value of $\langle B \rangle$ one comes as a result of thermolization to $\langle B \rangle = 0$ if baryonic charge is not conserved. This is not so if there are conserved combinations of charges including B. For example in $SU(5)$-models of grand unifications the difference $(B - L)$ is conserved (L is leptonic charge). Because of that $\langle B \rangle$ does not vanish even in the thermal equilibrium state if initially $\langle (B-L) \rangle_0 \neq 0$. The value of $\langle B \rangle$ in this case depends on the relation between the temperature and the masses of quarks and leptons. In particular for $T \gg m$, $\langle B_{eq} \rangle = \langle B - L \rangle_0 / 2$. As a result the baryon asymmetry depends on initial conditions. If however inflation preceded baryosynthesis, initial values of all conserved charges must be zero.

After we have explained the necessity of these three conditions for the generation of the asymmetry, let us note that the subsequent theoretical development showed that neither of them is obligatory. Baryogenesis may proceed with C and CP invariant theory in thermal equilibrium, and even if baryonic charge is strictly conserved in particle interactions. For the discussion of these counterexamples see the review paper by Dolgov (1992).

There are several qualitatively different scenarios of baryogenesis described in the literature starting from the classical one based on nonequilibrium decays of heavy particles, to more elaborate scenarios like squark condensate decay, or nonperturbative processes in electroweak interactions, and more exotic ones like concealment of conserved baryonic charge in evaporating black holes, or generation of baryonic excess due to CPT-violation or by external time-dependent field. The last ones are possible in thermal equilibrium.

It is noteworthy that inflationary model with a considerable expansion can be realized only if baryonic charge is not conserved. For successful solution of the flatness and horizon problems duration of inflationary stage should be sufficiently large, $H_I t_I \geq 65 - 70$. If baryonic charge were conserved it would be diluted during inflation by a huge factor $e^{210} - e^{195}$ as follows from the covariant law of the conservation of baryonic current. For the case of homogeneous and isotropic universe when space components of the current vanish it has the form

$$\partial_0 j_B^0 + 3H j_B^0 = 0 \qquad (0.2)$$

During inflationary stage, when $H = const$, baryonic charge density decreases as $j_B^0 \sim \exp(-3Ht)$ diminishing down to the value $3 \cdot 10^{-10}$. Unnatural by itself, the huge initial value of the baryonic charge density in principle may exist. But nonzero baryonic charge density implies simultaneously nonzero energy density associated with it. Inflation could be achieved only if energy density in the Universe is a constant or slowly varying function of the scale factor a. This is not true for the energy density associated with baryonic charge, ρ_B. It varies as $1/a^3$ for nonrelativistic particles and as $1/a^4$ for relativistic ones. ¿From the value of β (0.1) we may conclude that at high temperature stage $\rho_B \approx 10^{-10}\rho_{tot}$. It means that the total energy density could be approximately constant for the period not larger than 6 Hubble times which is too little for a successful inflation. Thus inflation demands nonconservation of baryons.

The idea of baryogenesis emerged from the observations that the Universe at some distance scale l_B around us is practically 100% charge asymmetric with baryon number density very much exceeding that of antibaryons, $N_B \gg N_{\bar{B}}$. The magnitude of the asymmetry is characterized by the ratio of the baryonic number density to the number density of photons in cosmic microwave background radiation (0.1). This small number means in particular that the size of the charge asymmetry (which is practically 100% now) was tiny at high temperatures, $T > \Lambda_{QCD} \approx 100$ MeV. At these temperatures antibaryons were practically equally abundant in the primeval plasma and correspondingly $(N_B - N_{\bar{B}}/(N_B + N_{\bar{B}}) \approx \beta \ll 1$. Still this number, though very small, is not easy to obtain and the main goal of theoretical models is to get this number as large as possible.

There are three important problems related to the scale of the asymmetry l_B:

- 1. What is the magnitude of l_B? Is it infinite or, what is practically the same, larger that the present-day horizon, $l_B > l_U \approx 10^{10}$ years? May it be rather small, say, like $a few \times 10$ Mpc? In the first case the whole Universe or at least the visible part is baryon dominated while in the second case there may be a considerable amount of antibaryons which can be in principle observed by their interaction with matter on the boundaries. Still since the distance is fairly large the gamma-flux from the annihilation would be sufficiently low.

- 2. May the Universe be charge asymmetric only in our neighbourhood, never mind how large it is (even larger than the horizon), and be charge symmetric as a whole? The last possibility is aesthetically appealing since particle-

antiparticle symmetry is restored on large.

- 3. Is the amplitude of the asymmetry β a constant or may it be a function of space points $\beta = \beta(x, y, z)$? The last case corresponds to the so called isocurvature density fluctuations which may be very interesting for the structure formation in low Ω Universe.

Historically first papers on baryogenesis which were based on a well defined particle physics model were done in the frameworks of the grand unification theories (for the review and the literature see Dolgov and Zeldovich (1981) and Kolb and Turner (1983)). Grand unification models present a beautiful extension of the minimal standard $SU(3) \times SU(2) \times U(1)$-model (MSM). A strong indication of the validity of the grand unification is the crossing of all three gauge coupling constants of supersymmetric extension of MSM at the same point near $E_{GUT} = 10^{16}$ GeV. On the other hand it is rather difficult to believe that there are no new particles and interactions in the region between electroweak or low energy supersymmetry scale and grand unification scale, but if the essential quantity is the logarithm of energy the distance between these two scales is not too big and one may hope that MSM or supersymmetric version of it is the ultimate truth in low energy physics (up to E_{GUT}). One more argument in favor of low energy supersymmetry is provided by cosmology, namely, if one demands in accordance with the theory of large scale structure formation that the bulk of matter in the universe is in the form of cold dark matter and assumes that the cross-section of the annihilation of the latter is given by $\sigma = \alpha^2/m^2$ then the mass m should be in the region 100 GeV - 1 TeV. It is just the scale of low energy supersymmetry.

A strong objection against GUT baryogenesis is a low heating temperature after inflation. It is typically 4-5 orders of magnitude below E_{GUT}. It means that the GUT era possibly did not exist in the early Universe. A very interesting alternative to the GUT baryogenesis is the electroweak one (for the review see Dolgov(1992) and Cohen, Kaplan, Nellson (1993)). Electroweak theory provides all the necessary ingredients for baryogenesis including baryon nonconservation (see below) so one may hope to get some baryon asymmetry of the Universe even in the frameworks of the MSM. A very interesting question is if it is possible to get the right magnitude of the asymmetry in MSM or baryogenesis demands an extension of the minimal model.

One may say in support of the second possibility that cosmology already demands physics beyond the standard model. It should be invoked for realization of

inflation, for the generation of the primordial density perturbations, for nonbaryonic dark matter, etc. A drastic change in the standard physics may be necessary for the solution of the cosmological term problem. (There is a hope however that it may be solved by infrared instability of quantum gravity in De Sitter background, see e.g. Ford(1985), Tsamis and Woodard(1993), Dolgov, Einhorn, and Zakharov (1994).) So we have already a strong evidence that there is physics beyond the standard model and thus baryogenesis should not be confined to the MSM. Still the possibility of realistic baryogenesis in the minimal model is extremely appealing and moreover it gives the unique possibility to express the magnitude of the baryon asymmetry β through parameters of the standard model measured in direct experiments.

Baryonic charge nonconservation in the electroweak theory was discovered by 't Hooft (1976). It is a very striking phenomenon. Classically baryonic current, as inferred from the electroweak Lagrangian, is conserved

$$\partial_\mu j^\mu_{baryonic} = 0, \tag{0.3}$$

but the conservation is destroyed by the quantum corrections. The latter are given by the very well known chiral anomaly associated with triangle fermionic loop in external gauge field. The calculation which can be found in many textbooks gives

$$\partial_\mu j^\mu_{BL} = N_f \left(\frac{g_2^2}{32\pi^2} W\tilde{W} - \frac{g_1^2}{32\pi^2} Y\tilde{Y} \right) \tag{0.4}$$

Here N_f is the number of fermionic flavors, $g_{1,2}$ are the gauge coupling constants of $U(1)$ and $SU(2)$ groups, W and Y are the gauge field strength tensors for $SU(2)$ and $U(1)$ respectively, and tilde means dual tensor, $\tilde{W}^{\mu\nu} = \epsilon^{\mu\nu\alpha\beta} W_{\alpha\beta}/2$. The products of the gauge field strength $W\tilde{W}$ and $Y\tilde{Y}$ can be written as divergences of vector quantities,

$$W\tilde{W} = \partial_\mu K_2^\mu \tag{0.5}$$

$$Y\tilde{Y} = \partial_\mu K_1^\mu \tag{0.6}$$

where

$$K_1^\mu = \epsilon^{\mu\nu\alpha\beta} Y_{\nu\alpha} Y_\beta \tag{0.7}$$

$$K_2^\mu = \epsilon^{\mu\nu\alpha\beta}(W_{\nu\alpha}W_\beta - \frac{1}{3}g_2 W_\nu W_\alpha W_\beta) \tag{0.8}$$

Here Y_ν and W_ν are gauge field potentials of abelian $U(1)$ and nonabelian $SU(2)$ groups respectively. Usually total derivatives are unobservable since they may be integrated by parts and disappear. This is true for the contribution into K^μ from the gauge field strength tensors $Y_{\mu\nu}$ and $W_{\mu\nu}$ which should sufficiently fast vanish at infinity. However it is not obligatory for the potentials for which the integral over infinitely separated hypersurface may be nonzero. Hence for nonabelian groups the current nonconservation induced by quantum effects becomes observable.

Because of conditions (0.4,0.5,0.8) the variation of the baryonic charge can be written as

$$\Delta B = N_f \Delta N_{CS} \tag{0.9}$$

where N_{CS} is the so-called Chern-Simons number characterizing topology in the gauge field space. It can be written as a space integral of the time component of the vector K^μ:

$$N_{CS} = \frac{g_2^2}{32\pi^2}\int d^3x K_2^i \tag{0.10}$$

Though N_{CS} is not a gauge invariant quantity its variation $\Delta N_{CS} = N_{CS}(t) - N_{CS}(0)$ is.

In vacuum the field strength tensor $W_{\mu\nu}$ should vanish while the potentials are not necessarily zero but can be the so called purely gauge potentials:

$$W_\mu = -\frac{i}{g_2}U(x)\partial_\mu U^{-1}(x) \tag{0.11}$$

There may be two classes of gauge transformations keeping $W_{\mu\nu} = 0$: one that does not change N_{CS} and the second that changes N_{CS}. The first one can be realized by a continues transformation of the potentials while the second cannot. If one tries to change N_{CS} by a continuous variation of the potentials one has to pass the region where $W_{\mu\nu}$ is nonzero. It means that vacuum states with different topological charges N_{CS} are separated by the potential barriers. The probability of the barrier

penetration can be calculated in quasiclassical approximation. The trajectory in the field space in imaginary time which connects two vacuum states differing by a unit topological charge is called the instanton. As in the usual quantum mechanics action evaluated on this trajectory gives the probability of the barrier penetration (Belavin et al (1975)):

$$\Gamma \sim \exp\left(\frac{4\pi}{\alpha_W}\right) \approx 10^{-170} \tag{0.12}$$

where $\alpha_W = g_2^2/4\pi$. This number is so small that it is not necessary to present a preexponential factor.

Expression (0.12) gives the probability of the baryonic charge violation at zero energy. We know from quantum mechanics that the probability of the barrier penetration rises with rising energy. Moreover in the system with nonzero temperature a particle may classically go over the barrier with the probability determined by the Boltzmann exponent, $\exp(-E/T)$. This analogy let one think that a similar phenomenon may exist in quantum field theory so that the processes with baryonic charge violation are not suppressed at high temperature. One should not of course rely very much on this analogy since there may be a serious difference between quantum mechanics which is a system with a finite number of degrees of freedom and quantum field theory which has an infinite (continuous) number of degrees of freedom. Still in a detailed investigation of this phenomenon convincing arguments have been found that baryonic charge nonconservation at high temperature may be strong and that baryogenesis by electroweak processes may be possible.

The first paper where this idea was seriously considered belongs to Kuzmin, Rubakov, and Shaposhnikov (1985) (for the earlier papers see review by Dolgov (1992)). They argued that the probability of baryonic charge nonconservation at nonzero T is determined by the expression

$$\Gamma \sim \exp\left(-\frac{U_{max}}{T}\right) \tag{0.13}$$

where U_{max} is the potential energy at the saddle point separating vacua with different topological charges. The field configuration corresponding to this saddle point is called sphaleron. It was originally found by Dashen, Hasslacher, and Neveu (1974) and later rediscovered by Manton (1983). In the last paper the relation of this solution to the topology changing transitions and baryonic charge nonconservation

was clearly understood. Quantum mechanical analogue of the sphaleron is a single point in the phase space, i.e. the position of particle sitting at the top of the barrier. The energy of the sphaleron is

$$U_{max} \equiv U(\phi_{sphaleron}(x)) = \frac{2M_W}{\alpha_W} f\left(\frac{\lambda}{g^2}\right) \tag{0.14}$$

where λ is the self-interaction coupling constant of the Higgs field, f is a function which can be calculated numerically, $f = O(1)$, and M_W is the mass of the W-boson. At zero temperature $2M_W/\alpha_W \approx 10$ TeV. However at high temperatures close to the electroweak phase transition the Higgs condensate is gradually destroyed and the height of the barrier decreases together with the mass of W-boson $M_W^2(T) = M_{0W}^2(1 - T^2/T_c^2)$ (see Kirzhnits, 1972 and Linde, 1979) where $T_c = O(1TeV - 100GeV)$ is the critical temperature of the transition. Thus one may expect that the processes with baryonic charge nonconservation are indeed unsuppressed at high temperatures.

The situation is not so simple however and there are a few problems which should be resolved before a definite conclusion can be made. They mostly stem from the difference between finite dimensional system like quantum mechanics and infinitely dimensional field theory. The first question is what is the probability of the processes with the change of topology in the gauge field space. Such processes proceed in presumably multiparticle collisions through formation of the classical field configuration with the coherent scale which is much larger than inverse temperature. If these processes are not fast enough the sphalerons may be not in thermal equilibrium and possibly far below the equilibrium so that the expression (0.13) would not be applicable. At the present day we do not know a reliable analytical way to address this problem. Numerical simulation of the analogous problem made in 1+1 dimensions by Grigoriev and Rubakov (1988) showed that the creation of soliton-antisoliton pairs are indeed fast enough to maintain the equilibrium value and this is one the strongest arguments in favor of efficient baryon nonconservation in electroweak processes. However such processes in one dimensional space may proceed much easier than those in three space dimensions simply because in $D = 1$ the change of topology means just a jump from one constant value of the Higgs field to another while in $D = 3$ much more fine tuning in every space point is necessary. Unfortunately numerical simulation in $3 + 1$ case is much more difficult and correspondingly much less reliable. So strictly speaking the probability of the sphaleron transitions is not known and a better understanding of it is very much desirable though it seems plausible that they are not too much suppressed so that thermal

equilibrium with respect to the topology changing transitions was achieved in the early universe.

Another question related to the probability of the processes with $\Delta B \neq 0$ is what is the entropy of the sphalerons or in other words what is the preexponential factor in expression (0.13). This factor characterizes the width of the potential near the saddle point in the directions orthogonal to the trajectory over potential barrier and was calculated by Arnold and McLerran (1987). With this factor taken into account the probability of electroweak processes with baryonic charge nonconservation in the phase with broken electroweak symmetry can be evaluated as

$$\frac{\Gamma_{\Delta B}}{H} = 10^{24} \left(\frac{M_W(T)}{T} \right)^2 e^{-120 M_W(T)/T} \tag{0.15}$$

where H is the Hubble parameter characterizing the rate of the Universe expansion.

At temperatures above electroweak phase transition the rate of baryonic charge nonconservation was given by Arnold and McLerran (1987) and by Khlebnikov and Shaposhnikov (1988):

$$\Gamma_{\Delta B} \approx \alpha_W^4 T \tag{0.16}$$

Recall that expressions (0.15) and (0.16) are valid only if sphalerons are in thermal equilibrium. If this is true then $\Gamma_{\Delta B}/H \gg 1$ at high temperatures and then abruptly falls down with falling temperatures. Thus processes with baryonic charge nonconservation are in equilibrium at high T and at some point are instantly switched off. Thus any preexisting baryon asymmetry would be washed out and a new one cannot be generated. This conclusion can be avoided however if deviations from thermal equilibrium existed at the time when baryonic charge nonconservation was still effective. This can be realized in particular if electroweak phase transition is of the first order. However it is still an open question what is the type of the phase transition depending in particular on the value of the Higgs boson mass.

One more comment may be in order here. We spoke before only about baryonic charge nonconservation. In fact electroweak interactions break equally baryonic and leptonic charges so that $(B - L)$ is conserved. With this correction in mind all the previous statements remain true with the substitution of $(B + L)$ instead of B.

Thus the following logical possibilities exist for the electroweak baryogenesis(we simply enumerate them here and discuss in some more detail giving recent references below):

- I. Change of the field topology is suppressed in three-dimensional space. Sphalerons are never abundant and electroweak nonconservation of $(B+L)$ is ineffective. In that case we should return either to GUT baryogenesis or to some other more recent proposals described in review paper by Dolgov (1992).

- II. Sphaleron transitions are not suppressed above and near the electroweak phase transition and so $(B+L)$ is strongly nonconserved at these temperatures. If this is true the following two possibilities are open:

- II.1. The electroweak phase transition is of the second order and so the baryon nonconserving processes, which were with a very good accuracy in thermal equilibrium above the phase transition, would be instantly completely switched off below it. In this case any preexisting $(B+L)$ would be washed out and we again meet two possibilities:

- 1a. The observed asymmetry might arise from an earlier generated $(B-L)$ either by $(B-L)$ nonconserved processes which exist e.g. in higher rank grand unification groups or by lepton charge nonconservation in decays of heavy Majorana fermion as was proposed by Fukugita and Yanagita (1986).

- 1b. Baryogenesis should take place at low energies below electroweak scale which for sure demands new low energy weak physics.

- II.2. Electroweak phase transition is first order so thermal equilibrium was strongly broken when both phases coexisted. If this is the case $(B+L)$-asymmetry could be generated in electroweak processes at temperatures near 1TeV. An important subdivision in this situation is:

- 2a. The standard model is able to give a correct magnitude of the baryon asymmetry of the Universe so that baryogenesis does not demand any physics beyond the minimal standard $SU(3) \times SU(2) \times U(1)$-model (MSM).

- 2b. An extension of the minimal standard model is necessary. This is not well defined and may include an introduction of additional Higgs fields (like in supersymmetric versions), considerable CP-violation in the lepton sector, CP-violation in strong interaction, etc.

The essential quantity which determines the character of the phase transition in the minimal standard model is the magnitude of the Higgs boson mass. For a large value of the latter the phase transition is second order and for a small one it is first order. To illustrate this statement let us consider the following temperature dependent effective potential for the Higgs field ϕ (temperature dependent terms appear due to interactions of the field ϕ with the thermal environment of the cosmic plasma):

$$U(\phi,T) = m^2(T)\phi^2/2 + (\lambda\phi^4)\ln(\phi^2/\sigma^2)/4 + \gamma(T)\phi^3 + ... \qquad (0.17)$$

Notations here are selfexplanatory. The temperature dependence of the effective mass is roughly speaking the following $m^2(T) = -m_0^2 + AT^2$ where the constant A is usually positive. (It is positive in MSM.) Logarithmic dependence on ϕ came from one-loop quantum perturbative corrections to the potential. At high temperatures the potential has the only minimum at $T = 0$, vacuum expectation value of the ϕ is zero, and the electroweak symmetry is unbroken. At smaller temperatures a deeper minimum is developed at nonzero ϕ and mass of the field near this minimum is $m_H^2 \approx 2m_0^2$ (we neglected here logarithmic terms in U). One sees that the larger is m_0^2 (and correspondingly the physical mass m_H^2) the easier is the phase transition. There is no consensus in the literature about the value of m_H separating first and second order phase transitions. While earlier perturbative calculations in the MSM by Shaposhnikov (1987) give a rather small value $m_H \approx 45$ GeV, it was argued that higher loop effects are essential (Dine et al (1992), Arnold and Espinosa (1993), Barnasco and Dine (1993)). Moreover since thermal perturbation theory for nonabelian gauge fields suffers from severe infrared divergences (see Linde (1979)), nonperturbative effects might be important acting in favor of the first order phase transition with higher m_H (Shaposhnikov (1993)). For a more detailed discussion and list of references see papers by Cohen, Kaplan, and Nelson (1993) and by Farrar and Shaposhnikov (1993a). Hence we cannot make any rigorous conclusion now about the nature of the electroweak phase transition though it seems probable that MSM with the existing lower experimental bound on the Higgs mass $m_H > 62$ GeV given by LEP favors second order phase transition while in extended models with several Higgs fields the transition might be first order.

Even if the electroweak phase transition in MSM is first order the generated asymmetry is expected to be very small. It is connected with a strong suppression of CP-violating effects at high temperatures. CP-breaking in the MSM is created by the imaginary part of the quark mass matrix (Cabibbo-Kobayashi-Maskawa matrix). If there are only two quark generations the imaginary part is not observable because the

phase may be absorbed in a redefinition of the quark wave function. The statement remains true with more quarks families with degenerate masses because the unit matrix is invariant with respect to unitary transformations. One can see that the minimum number of quark families for which the imaginary part is observable is three with different masses of quarks with the same value of electric charge. (If we believe that there is no extension of the standard model then the necessity of CP-violation for the generation of the charge asymmetry of the Universe justifies the existence of at least three fermionic families.) Moreover the amplitude of CP-violation is proportional to the mixing angles between different families because if the quark mass matrix and the kinetic term in the Lagrangian are simultaneously diagonal then the phase rotation would not change them. By these reasons the amplitude of CP-violation in MSM is suppressed by the factor (which is called the Jarlskog determinant):

$$A_- \sim \sin\theta_{12}\sin\theta_{23}\sin\theta_{31}\sin\delta_{CP}(m_t^2-m_u^2)(m_t^2-m_c^2)$$
$$(m_c^2-m_u^2)(m_b^2-m_s^2)(m_b^2-m_d^2)(m_s^2-m_d^2)/E^{12} \qquad (0.18)$$

Here θ_{ij} are mixing angles between different generations and δ_{CP} is the CP-odd phase in the mass matrix. The product of sin's of these quantities is about $10^{-4}-10^{-5}$. E is the characteristic energy of a process with CP-breaking. In the case considered when the temperature of the medium is above 100 GeV, E is of the same order of magnitude. Correspondingly one should expect that baryon asymmetry in MSM should be of the order of 10^{-20}.

This conclusion was questioned recently by Farrar and Shaposhnikov (1993a,b). They argued that flavor dependent temperature corrections to the quark masses in the vicinity of the domain wall where the expectation value of the Higgs field is changing nonadiabatically, may drastically enhance efficiency of the electroweak baryogenesis. This effect is especially pronounced at the low energy tail of the quark distribution in the phase space. As a result the value of the baryon asymmetry may be close to the observed one even in the minimal standard model. This very interesting proposal was however strongly criticized recently.

Despite all the attractiveness of the possibility of effective baryogenesis in the MSM it should be excluded if the experimental lower bound on the Higgs boson mass proves to be above the value necessary for successful first order phase transition. This seems rather probable now and the models with several Higgs fields are possibly the next best choice. They may give a larger CP-violation and what's more in these models both experimental and theoretical bounds on the Higgs boson mass are much

less restrictive.

The generic feature of all scenarios of electroweak baryogenesis is a coexistence of two phases in one of which baryonic charge is strongly nonconserved, the corresponding reactions are well in equilibrium, and no asymmetry can be generated, while in the second phase baryonic charge is practically conserved and the asymmetry also cannot be generated though by an opposite reason. So the only place where baryon asymmetry may be produced are the boundaries between the phases. The outcome of such a process strongly depends upon the interaction between the high temperature cosmic plasma and the domain walls and in particular upon the velocity of the wall propagation in plasma. These problems are addressed in several papers (for the recent ones see e.g. Liu, McLerran, and Turok (1993) and Huet et al (1993)) but still more work in this field is desirable.

In the case if the phase transition is second order, baryon asymmetry could not be generated by electroweak processes but, if sphalerons are effective, the latter may be very good for erasure of any preexisting $(B+L)$-asymmetry. A nonzero initial $(B-L)_i$-asymmetry is conserved by electroweak interactions and the subsequent sphaleron processes would result in equal baryon and lepton asymmetry $B_f = L_f = (B-L)_i/2$. Assuming that this is indeed the case one can derive a bound on the strength of $(B-L)$-nonconserving interactions at lower temperatures when (and if) $(B+L)$-erasure is effective. (One should keep in mind however that all these bounds are valid only if there is no baryogenesis at electroweak or lower temperature range.) If the rate of $(B+L)$-nonconserving sphaleron transitions is given by eqs.(0.15, 0.16), the sphaleron processes are in equilibrium in the temperature range

$$10^2 - 10^3 < T < 10^{12}\,(GeV) \tag{0.19}$$

For successful baryogenesis the processes with $(B-L)$-nonconservation should not be in equilibrium in this range. This idea was first used by Fukugita and Yanagita (1986, 1990), who proposed the model of baryogenesis through the decay of heavy Majorana fermion, to put a bound on the Majorana mass of light neutrinos, $m_M(\nu) < 50$ KeV. Neutrinos with a larger Majorana mass together with sphalerons would destroy both baryon and lepton asymmetry. The assumption that baryogenesis proceeds through transformation of an initial $(B-L)$-asymmetry into B-asymmetry permits to deduce in some cases more interesting bounds on e.g. L-nonconservation than that following from direct experiments. There are too many possible forms of the interaction and theoretical models giving rise to them so that their more detailed description is outside the scope of the present talk and one should be addressed to

original literature on the subject. In this respect the lectures by Olive (1994) on big-bang baryogenesis would be very helpful.

Now I would like to turn to some more exotic cases. The first one is a possibility of a large lepton asymmetry together with a normal small baryon asymmetry. Though the data gives a rather accurate value of β (within an order of magnitude), the value of the lepton asymmetry is practically unknown. The best limits follow from the primordial nucleosynthesis which permits muonic and taonic lepton asymmetry close to unity while electronic lepton asymmetry cannot exceed 1% (see Dolgov (1992) for the list of references). The bound on the chemical potential associated with electronic charge is stronger because it would directly shift proton-neutron equilibrium in weak reactions like $n + \nu_e \leftrightarrow p + e^-$, while ν_μ and ν_τ influence n/p-ratio only through the total energy density. Thus even in the most restricted case the value of lepton asymmetry may be as large as 10^{-2}.

A large lepton asymmetry could only be realized if the sphaleron processes were not effective or if the asymmetry was generated below electroweak scale. Even if this is true, the majority of models naturally give $L \approx B$ but there are some examples permitting $L \gg B$ (see Dolgov and Kirilova (1991) and Dolgov (1992)). In this case we would have at our disposal an extra free parameter for the theory of primordial nucleosynthesis, namely the chemical potential of leptons. What's more the characteristic scale of spatial variation of the leptonic charge density l_L might be much smaller than l_B and if the former is in the range $l_{gal} < l_L < l_U$ one may observe that by spatial variation of the abundances of light nuclei and in particular of ^4He.

The relatively strong isocurvature fluctuations in leptonic sector with a possibly nonflat spectrum may be also interesting for the theory of the large scale structure formation with a single dominant component of hot dark matter. Usually one considers isocurvature perturbations in baryonic sector which are stronger bounded by the isotropy of the cosmic microwave background.

Returning to the isocurvature fluctuations in baryonic sector one may find plenty baryogenesis scenarios (see Dolgov (1992)) providing very interesting perturbations with the spectrum varying from the flat one to that having a prominent peak at a particular wave length. The last case corresponds to a periodic in space distribution of baryonic matter. It may be naturally realized if three rather innocent assumptions are satisfied:

- 1.There exists a complex scalar field ϕ with the mass which is small in comparison with the Hubble parameter during inflation. The latter may be as large as 10^{14} GeV so one does not need a really light scalar field.
- 2. The potential of the field ϕ contains nonharmonic terms like $\lambda|\phi|^4$.
- 3.A condensate of ϕ was formed during inflationary stage which was a slowly varying function of space points.

If these conditions are fulfilled then it can be proven (for the details that the distribution of baryons in the Universe would be in the form:

$$N_B \approx N_{B0} + N_1 \cos \frac{\vec{r}\vec{n}}{l_B} \tag{0.20}$$

where $\vec{n}$ is an arbitrary unit vector. The scale l_B of the fluctuations is given by the exponentially stretched Compton wave length of ϕ and could easily be as large as 100 Mpc as was observed by Broadhurst et al (1990). An interesting picture emerges if $N_0 = 0$ and the Universe consists of alternating baryonic and antibaryonic layers.

Another unusual picture of the Universe, the so called island universe model may be realized with the specific though not too complicated model of baryogenesis (Dolgov et al 1987). In this model our Universe is a huge baryonic island with the size large or about 10^{10} years (or $z = 5 - 10$), while floating in the see of dark matter which is more or less uniformly distributed. There are two interesting features of this model which may be relevant to the structure formation. First, the background radiation comes to us from the baryon empty regions so that the fluctuations in its temperature is not directly related to the density perturbations inside the island. Second, our noncentral position inside the island would give rise to intrinsic dipole, $d \sim 10^{-3}$, in the angular distribution of the microwave radiation which is not related to our motion. The quadrupole asymmetry in this case would be rather small, $q \sim d^2 \sim 10^{-6}$. It may make easier structure formation in the cold dark matter model. Without intrinsic dipole and with the flat spectrum of perturbations more complicated models of the structure formation are necessary, like e.g. a mixture of hot and cold dark matter or a model with cold dark matter and nonzero vacuum energy (cosmological constant). Both these models demand some fine tuning which is not well understood today. The first one needs the energy density of hot and cold dark matter to be the same within the factor of 2 while the other demands ρ_{vac} which is normally time independent constant to be close

today to the critical energy density which is time dependent, $\rho_c \sim m_{Pl}^2/t^2$. The latter may be explained if the smallness of the cosmological constant is ensured by the so called adjustment mechanism (for the review see Weinberg (1989) or Dolgov (1989)). Though these two possibilities are more conservative than the island model still they are not the most economic ones. Proliferation of the universe components from the purely baryonic universe to the mixed baryonic and hot dark matter or later on to baryonic and cold dark matter and now to the mixture of all three of them (baryonic+cold+hot) with close energy densities is rather mysterious. On the other hand there are stable neutrinos which are very likely to be massive and it is also very plausible that there is supersymmetry in particle physics so that there should be a stable heavy particle. These two are perfect candidates for the hot and cold dark matter (what's more we may have now dark solar size objects in galaxies) so that it would be only natural that these particles participates as building blocks of the Universe. The unresolved question is their interaction strength which provides very different number densities and similar mass densities for the particles of hot and cold dark matter.

One may try to make a cosmological model assuming that the only massive stable particles in the Universe are protons and electrons and all the dark matter is made of the normal baryonic staff. To do that one has to develop a scenario in which baryogenesis proceed much more efficiently in relatively small space regions giving $\beta = 1 - 0.01$ while it goes normally outside (Dolgov and Silk (1993)). The regions with that huge baryon number density mostly form black holes with the mass distribution

$$\frac{dN}{dM} \sim \exp\left(-\gamma \ln^2 \frac{M}{m_0}\right) \tag{0.21}$$

Parameters γ and M_0 cannot reliably found in the model but one reasonably expect that $\gamma = O(1)$ and M_0 is close to the solar mass. These black holes might be the objects observed in the microlensing observations reported here. If there are no other massive stable particles one has to build a theory of the structure formation with these black holes which behave as normal cold dark matter. At the tail of the distribution in mass there should be very heavy black holes with masses like $10^6 - 10^9$ solar masses which may serve as seeds for the structure formation. Still tilted spectrum of the initial perturbations may be desirable if only cold dark matter is permitted.

Conclusions to this lecture reflect to a large extend my personal opinion and may not be shared by everybody or not even by the majority.

- I. The best choice for the baryogenesis scenario is the electroweak one and in its framework the one based on the minimal standard model is the most appealing. The problems with the electroweak baryogenesis are the unknown probabilities of three dimensional reactions with classical field configurations, which may question the scenario as a whole, and the type of the electroweak phase transition. The knowledge of the value of the Higgs boson mass could be of great help here.

- II. If not MSM the low energy SUSY is the next best choice.

- III. If electroweak interactions destroy but not generate baryon asymmetry (like e.g. in the case of the second order phase transition), a very interesting possibility is baryogenesis through leptogenesis. One needs to this end a heavy Majorana fermion with mass around 10^{12} Gev (plus-minus a few orders of magnitude) and correspondingly a new physics beyond the standard model.

- IV. A very low temperature (below the electroweak scale) baryogenesis is not excluded but there is no natural particle physics model for that.

- V. Majority of models give lepton and baryon asymmetry of approximately the same magnitude but one may find scenarios giving $L \gg B$ with interesting consequences for the primordial nucleosynthesis.

- VI. A better understanding of baryogenesis may be of interest for the theory of the large scale structure formation in particular because in the process of baryogenesis isocurvature density fluctuations with a complicated spectrum might be created.

Chapter 5

Thermodynamics of elementary particles in the expanding Universe

The Universe expansion evidently means that the matter density was higher in the past. Higher were also the energies of separate particles because their momenta are redshifted inversely proportionally to the scale factor $p \sim a^{-1}$ (the wave length of a free particle rises with all the distances as the scale factor, $\lambda \sim a$). Thus the closer to the beginning was the Universe the higher and denser was the matter and the higher were the energy of interacting particles in the primeval plasma. Studying the history of the early Universe one can get important information about the properties of elementary particles.

The derivation of the cosmological bounds on the particle properties is based to the large extent on the remarkable fact that the matter in early the Universe was mostly in the state of thermal equilibrium. Of course there were essentially nonequilibrium stages as e.g. inflation or periods when some particular forms of matter were out of equilibrium but typical state was thermal equilibrium. In usual thermodynamics of stationary systems the longer is the time which passed from the beginning the better equilibrium is established. Cosmological case is just the opposite. Indeed the approach to equilibrium is determined by the relation between the characteristic reaction rates Γ and the expansion rate $H = \dot{a}/a$. Though for the Friedman expansion law $H \sim t^{-1}$ the reaction rate Γ increases even faster when

approaching the beginning, $t \to 0$. Indeed for Γ the following estimate is valid

$$\Gamma \equiv \dot{N}/N \approx \sigma N \tag{0.1}$$

where N is the particle number density and σ is the characteristic cross section of their interactions. Since on the relativistic stage $N \sim a^{-3} \sim t^{-3/2}$ the reaction rate exceeds H if the cross section does not decrease with the rising particle energies. The latter is not true at very high energies but as we see in what follows there is a considerable interval of time when the equilibrium condition

$$\Gamma > H \tag{0.2}$$

is satisfied. This is connected with the large value of the Planck mass in comparison with the characteristic mass parameters in particle physics.

Thermodynamic equilibrium permits to introduce the notion of the particle temperature T. For sufficiently weak interaction between particles the distribution of particles in momentum is given by the known Fermi or Bose-Einstein formulae for the ideal gas (see e.g. Landau and Lifshits, 1964):

$$n_{f,b}(p) = \{\exp[(E-\mu)/T] \pm 1\}^{-1} \tag{0.3}$$

Here signs '+' and '−' refers to fermions and bosons respectively, $E = \sqrt{p^2 + m^2}$ is the particle energy, and μ is their chemical potential. As is well known, chemical potentials of particles and antiparticles in equilibrium are equal in absolute values but opposite in signs:

$$\mu + \bar{\mu} = 0 \tag{0.4}$$

This follows from the equilibrium condition for the chemical potentials which for an arbitrary reaction $a_1 + a_2 + a_3 \ldots \to b_1 + b_2 + \ldots$ has the form

$$\sum_i \mu_{a_i} = \sum_j \mu_{b_j} \tag{0.5}$$

(see below the discussion of kinetic equation) and from the fact that particle and antiparticle can annihilate into different number of photons, $a + \bar{a} \to 2\gamma, 3\gamma, \ldots$. In particular chemical potential of photons vanishes in equilibrium.

One can see from eq.(0.3) that chemical potential of bosons cannot exceed their mass. On the other hand we believe that charge asymmetry may be arbitrarily large. How this may be compatible with the statement that it is proportional to the value of the chemical potential? We may put this problem in a more impressive form: assume that there is a gas of massless bosons. Can there be a charge asymmetry in this gas? The answer is of course "yes" but the result is not proportional to the value of the chemical potential which should be zero in this case. To ensure the asymmetry the distribution of bosons should be of the form:

$$n_b = (\exp[(E-m)/T]-1)^{-1} + C\delta^3(\vec{p}) \tag{0.6}$$

Note that we substituted m for the magnitude of the chemical potential. One can verify that this distribution function is indeed a stationary solution of the kinetic equation and the arbitrary constant C gives the magnitude of the asymmetry. Its magnitude is different for particles and antiparticles. This distribution corresponds to the formation of Bose condensate.

If certain particles possess a conserved charge their chemical potential in equilibrium can be nonvanishing. It corresponds to nonzero density of this charge in plasma. Thus plasma in equilibrium is completely described by temperature and by a set of chemical potentials corresponding to all the conserved charges. It follows from the observations that the densities of all charges in the Universe that can be measured is very small or even zero. So in what follows we will usually assume that $\mu = 0$. An exception is the discussion of the baryon asymmetry of the Universe when nonzero though very small μ_B is generated on nonequilibrium stage.

The number density of bosons corresponding to distribution (0.3) at $\mu = 0$ is

$$N_b \equiv \sum \int \frac{n_b(p)}{(2\pi)^3} d^3p = \begin{cases} \zeta(3) g T^3/\pi^2 \approx 0.12 g T^3, & \text{if } T > m; \\ (2\pi)^{-3/2} g (mT)^{3/2} exp(-m/T), & \text{if } T < m. \end{cases} \tag{0.7}$$

Here $\zeta(3) \approx 1.2$ and g is is the number of spin states of the boson. It comes from the summation over bosonic polarization states. In particular the number density of photons is

$$N_\gamma = 0.24 T^3 = 394 (T/2.7K)^3 cm^{-3} \tag{0.8}$$

For fermions the equilibrium number density is

$$N_f = \begin{cases} \frac{3}{4}N_b \approx 0.09T^3, & \text{if } T > m; \\ N_b \approx (2\pi)^{-3/2} g (mT)^{3/2} exp(-m/T), & \text{if } T < m. \end{cases} \tag{0.9}$$

The energy density of particles in plasma is equal to

$$\rho = \sum \frac{1}{2\pi^2} \int \frac{dp p^2 E}{exp(E/T) \pm 1} \tag{0.10}$$

Here the summation is done over all particle species in the plasma and their spin states. In the relativistic case

$$\rho = (\pi^2/30) K T^4 \tag{0.11}$$

where $K = \sum [g_b + (7/8) g_f]$. In particular for photons we get

$$\rho_\gamma = \frac{\pi^2}{15} T^4 \approx 0.25 \left(\frac{T}{2.7K}\right)^4 \frac{eV}{cm^3} \approx 4 \cdot 10^{-13} \left(\frac{T}{2.7K}\right)^4 \frac{erg}{cm^3} \tag{0.12}$$

The contribution of heavy particles, i.e. with $m > T$, into ρ is exponentially small if the particles are in thermodynamic equilibrium:

$$\rho(m > T) = gm \left(\frac{mT}{2\pi}\right)^{3/2} \exp\left(-\frac{m}{T}\right) \left(1 + \frac{27T}{8m} + \ldots\right) \tag{0.13}$$

As we see below the equilibrium for stable particles sooner or later breaks because their number density becomes too small to maintain the proper annihilation rate. Hence their number density drops as a^{-3} and not exponentially. This ultimately results in massive particle dominance in the Universe. Their number density is even larger if they possess a conserved charge and the corresponding chemical potential is nonvanishing. Unstable particles always maintain the equilibrium with the products of their decay because the decay rate Γ does not depend on the particle number density and the expansion rate goes down so at sufficiently large t the condition $\Gamma > H$ is fulfilled.

Since Ω is very close to unity on the early stage the energy density at that time almost coincides with the closure density (0.10). Taking this into account it is

easy to find the dependence of temperature on time on RD-stage when $H = 1/2t$ and ρ is given by eq. (0.11):

$$T^2 = \left(\frac{90}{32\pi^3}\right)^{1/2} \frac{m_{Pl}}{\sqrt{Kt}} = \frac{2.42\,(T/Mev)^2}{\sqrt{K}\;\;(t/sec)} \tag{0.14}$$

In the course of expansion and cooling down K decreases as long as the particle species with $m > T$ disappear from the plasma.

If all the chemical potentials vanish and thermal equilibrium is established the entropy of the primeval plasma is conserved

$$\frac{d}{dt}\left(a^3 \frac{p+\rho}{T}\right) = 0 \tag{0.15}$$

Indeed from the well known relation

$$dE = d(\rho V) = TdS - pdV + \mu dN \tag{0.16}$$

it follows for $\mu = 0$ that

$$dS(V,T) \equiv d\left(V\frac{p+\rho}{T}\right) = \frac{1}{T}[d(\rho V) + pdV] \tag{0.17}$$

Since for $\mu = 0$ the equilibrium values of ρ and p are functions of temperature only, $\rho = \rho(T)$ and $p = p(T)$ it follows from eq. (0.17) that

$$\frac{\partial S}{\partial V} = \frac{p+\rho}{T} \tag{0.18}$$

$$\frac{\partial S}{\partial T} = \frac{V\partial\rho}{T\partial T} \tag{0.19}$$

Using the evident relation $\partial^2 S/\partial V\partial T = \partial^2 S/\partial T\partial V$ we get

$$dp = (\rho + p)dT/T \tag{0.20}$$

Conservation law (0.15) follows from this equation and eq. (0.8).

Now let us consider the variation of the particle number density in the expanding Universe. In the Boltzman kinetic equation the account should be taken of external gravitational field. In the homogeneous and isotropic case when the phase space distribution function for particle i depends only its momentum p_i and time t the extra term in the equation is particularly simple:

$$\frac{dn_i}{dt} = \frac{\partial n_i}{\partial t} + \frac{\partial n_i}{\partial p_i}\dot{p}_i = \frac{\partial n_i}{\partial t} - Hp_i\frac{\partial n_i}{\partial p_i} \tag{0.21}$$

where we have taken into account the redshifting of the momentum, $\dot{p} = -Hp$. As a result the kinetic equation takes the form

$$\left(\frac{\partial}{\partial t} - Hp_i\frac{\partial}{\partial p_i}\right) n_i(p_i, t) = S_i \tag{0.22}$$

where S_i is the collision integral:

$$S_i = -\frac{(2\pi)^4}{2E_i}\sum_{Z,Y}\int d\nu_Z\, d\nu_Y\, \delta^4(p_i + p_Y - p_Z) \tag{0.23}$$

$$\left[\mid A(i + Y \to Z)\mid^2 n_i \prod_Y n \prod_Z (1 \pm n) - \mid A(Z \to i + Y)\mid^2 \prod_Z n \prod_{i+Y}(1 \pm n)\right] \tag{0.24}$$

Here Y and Z are arbitrary generally multiparticle states, $\prod_Y n$ is the product of number densities of particles forming the state Y, and

$$d\nu_Y = \prod_Y \frac{d^3p}{(2\pi)^3 2E} \tag{0.25}$$

The sign '+' or '−' in $\prod(1 \pm n)$ is chosen for bosons and fermions respectively.

It can be easily checked up that in the stationary case ($H = 0$) distribution (0.3) with the account of conservation of energy $E_i + \sum_Y E = \sum_Z E$, and chemical potential, $\mu_i + \sum_Y \mu = \sum_Z \mu$ (0.5) are indeed the solutions of kinetic equation (0.22). This follows from the validity of the relation

$$\prod_{i+Y} n \prod_Z (1 \pm n) = \prod_Z n \prod_{i+Y} (1 \pm n) \tag{0.26}$$

and from the detailed balance condition, $\mid A(i + Y \to Z) \mid = \mid A(Z \to i + Y) \mid$. This condition is true if the theory is invariant with respect to time reversion. It is established in experiment however that this invariance (T-invariance) is only approximate. To be more exact it is known that CP-invariance is broken and hence because of CPT-theorem T-invariance is also broken. Still even if the detailed balance condition is violated the form of the equilibrium distribution functions remain the same. This is ensured by the weaker condition

$$\sum_k \int d\nu_{Z_k} \delta^4 \left(\sum_{Z_k} p - p_f \right) \left(\mid A(Z_k \to f) \mid^2 - \mid A(f \to Z_k \mid^2 \right) = 0 \qquad (0.27)$$

which can be called the cyclic balance condition. It follows from the unitarity of S-matrix, $S^+S = SS^+ = 1$. In fact a weaker condition is sufficient for saving the standard form of the equilibrium distribution functions, namely the diagonal part of the unitarity relation, $\sum_f W_{if} = 1$ and the inverse relation $\sum_i W_{if} = 1$ where W_{if} is the probability of transition from the state i to the state f. The condition that the sum of probabilities of all possible events is unity is of course trivial. Slightly less trivial is the inverse relation which can be obtained from the first one by the CPT-theorem.

Eq. (0.22) with collision integral (0.24) is valid in the ideal gas approximation when the particle interactions are weak enough. The criterion of its applicability is the large size of the particle mean free path $l_f = (\sigma N)^{-1}$ in comparison with the average distance between the particles, $l_T = N^{1/3} \sim T^{-1}$. This condition is true e.g. in gauge theories for which $\sigma \sim \alpha^2/T^2$ if $T > m_V$ and $\sigma \sim \alpha^2 T^2/m_V^4$ if $T < m_V$ where m_V is the mass of the intermediate boson and the coupling constant $\alpha \ll 1$.

We will consider first the case when the particle interactions can be neglected and the simple equation is valid

$$\left(H^{-1} \frac{\partial}{\partial p} - p \frac{\partial}{\partial p} \right) n(p,t) = 0 \qquad (0.28)$$

Integrating this equation over d^3p and making the natural assumption that $n \to 0$ when $p \to \infty$ we get

$$\dot{N} = -3HN \qquad (0.29)$$

where $N = \int d^3pn(p,t)/(2\pi)^3$ is the particle number density. We have obtained the natural result that the number density of noninteracting particles decreases in the course of expansion as a^{-3}.

It is convenient to introduce the dimensionless variables

$$y = m_0/T \tag{0.30}$$

$$\lambda_i = p_i/T \tag{0.31}$$

where m_0 is a parameter with dimension of mass and the temperature T is formally defined by the condition

$$\dot{T} = -HT \tag{0.32}$$

Defined in this way the variable T can be used even in nonequilibrium case.

In terms of variables y and λ eq. (0.28) has the very simple form

$$y\frac{\partial n}{\partial y} = 0 \tag{0.33}$$

Hence for free particles $n(p,t)$ is the function of the ratio $\lambda = p/T$ only. In particular for massless particles the equilibrium form of the distribution function is maintained in the course of expansion even when the interaction is switched off. The well known example of this is the spectral distribution of the cosmic microwave background radiation. For massive particles initially (at $T = T_0$) equilibrium distribution function $f(E/T)$ goes into $f(\sqrt{m^2/T_0^2 + p^2/T^2})$. In the last expression the parameter T is not of course the temperature of these particles.

In the course of the Universe expansion the interactions between particles in the primeval plasma are effectively switched off and the number density of stable particles x in the comoving volume or in other words the ratio N_x/N_γ tends to a constant value (see eq. (0.29)). This phenomenon is called concentration quenching in Russian literature and freezing in the English one. The limiting value of the product $N_x a^3$ depends on the strength of x-particle interactions. The stronger is the interaction the longer are the particles in the thermal equilibrium with the plasma

and the lower their limiting concentration. Roughly speaking $N_x/N_\gamma \sim exp(-m/T_f)$ where T_f is the temperature at which the interactions are switched off.

Two cases should be distinguished when the frozen value of the number density is calculated. The first one is $T_f > m_x$ and at the moment of decoupling $N_x/N_\gamma = O(1)$. The second case is $T_f < m_x$ and $N_x/N_\gamma \ll 1$ at the decoupling. The first case is realized for light neutrinos. The cross-section of their annihilation into e^+e^--pairs for $E_\nu > m_e$ is

$$\sigma(\nu_k\bar{\nu}_k \to e^+e^-) = 2C_k G_F^2 s/3\pi \tag{0.34}$$

where $s = (p_1 + p_2)^2$ is the total energy squared in the center of mass system, $G_F = 10^{-5}m_N^{-2}$ is the weak interaction coupling constant and the constant coefficients C_k depend upon the neutrino type, $C_e = \sin^4\theta_W + \left(\sin^2\theta_W + 1/2\right)^2 \approx 5/8$, $C_\mu = C_\tau = \sin^4\theta_W + \left(\sin^2\theta_W - 1/2\right)^2 \approx 1/8$. The annihilation cross-sections of ν_e and $\nu_{\mu,\tau}$ are different because at $T < m_\mu$ ν_μ and ν_τ interact only with neutral current while ν_e interacts both with neutral and charged currents. The neutrino decoupling temperature is approximately determined by the equality of the annihilation rate σN_ν and the expansion rate $H = 1/2t$. For the estimate of the decoupling temperature let us substitute into $s = 4E^2$ the average energy of neutrino $E \approx 3T$ and express H through T by eq. (reft2). Thus we get

$$T_{\nu f} \approx 2.7C_k^{-1/3}MeV \approx \begin{cases} 3Mev & \text{for } \nu_e; \\ 5MeV & \text{for } \nu_\mu \text{ and } \nu_\tau \end{cases} \tag{0.35}$$

In this calculation we put $K = 10.75$. This accounts for the contribution of photons ($K_\gamma = 2$), e^+e^--pairs ($K_e = 7/2$) and three species of left-handed neutrinos ($K_\nu = 21/4$).

If neutrinos are stable then as we see in what follows their mass should be substantially smaller then $T_{\nu f}$. (For ν_e it is known from direct experiments.) Hence at the moment of decoupling $N_{\bar{\nu}f} + N_{\nu f} = 3N_\gamma/4$. To the present time this ratio becomes somewhat smaller because the photon number density has increased as a result of e^+e^--annihilation at $T < m_e$. The increase of N_γ can be calculated if the entropy conservation law (0.15) is used. Comparing the effective number of degrees of freedom in the primeval plasma before and after the annihilation we find that the relative photon concentration increases 11/4 times. Hence the number density of each neutrino type at the present time is

$$N_{\nu 0} + N_{\bar{\nu} 0} = \frac{3}{11} N_{\gamma 0} \approx 110 \left(\frac{T_{\gamma 0}}{2.7K} \right)^3 cm^{-3} \tag{0.36}$$

Note that it has been implicitly assumed that there is no other sources of photons except for e^+e^--annihilation.

The calculated limiting value of the number density and the data on the Universe age t_U permit to get the upper bound on the neutrino mass (Gerstein and Zel'dovich, 1966). Using to this end expression (0.16) we get

$$\sum m_{\nu_i} < 380eV \left(\frac{2.7K}{T_\gamma} \right)^3 \left(\frac{0.98 \cdot 10^{10} y}{t_U} - h_{100} \right)^2 \tag{0.37}$$

Here the Universe age t_U should be larger than $0.98 \cdot 10^{10} y / h_{100}$. For large t_U and/or h_{100} this expression gives much stronger bound than the one usually presented in the literature. The older is the Universe the stronger is the upper limit on m_ν.

Inflationary universe models predict $\Omega_{tot} = 1$. If this is true the bound on m_ν depends only on H because for $\Omega = 1$ $t_U = 2/3H$. In this case it reads

$$\sum m_{\nu_i} < 95eV \left(\frac{2.7K}{T_\gamma} \right)^3 h_{100}^2 \approx 40eV \left(\frac{2.7K}{T_\gamma} \right)^3 \left(\frac{10^{10} y}{t_U} \right)^2 \tag{0.38}$$

The best upper bound is obtained from the minimum of these two.

Though the accuracy in determination of t_U and H is very low the existing trend towards large values shows some inconsistency between them. This makes one think about possible modifications of the standard expansion scenario. The simplest possibility is a nonzero cosmological constant Λ. In this case the Universe age is given by eq. (43) with $\Omega_r = \Omega_s = \Omega_w = 0$. If we assume once again that the inflationary model is valid and so $\Omega = 1$ the integral (0.43) is explicitly calculated

$$t_U = \frac{6.5 \cdot 10^9 y}{h_100} \Omega_v^{-1/2} ln \frac{1 + \sqrt{\Omega_v}}{1 - \sqrt{\Omega_v}} \tag{0.39}$$

If $\Omega_v > 0$ the values $h_{100} = 1$ and $t_U > 1.5 \cdot 10^{10}$years are compatible but the upper bound on m_ν becomes stronger than for $\Omega_v = 0$ (Dolgov, 1984):

$$\sum m_{\nu_i} < 95eV(2.7K/T_\gamma)^3(1-\Omega_v)^2 h_{100}^2 \qquad (0.40)$$

If $\Omega_v < 0$ the trend to contradiction between the large values of y_U and H becomes stronger but the restriction on m_ν is weaker. In particular for $t_U > 10^{10}$years and $h_{100} \geq 0.5$ one gets $\sum m_{\nu_i} < 45$eV.

The absence of a reliable model explaining the small value of Λ-term makes the bound on m_ν uncertain within the factor of 2 or 3 because there could be modifications of the expansion regime connected with the adjusting of Λ to zero which are not taken into account in the standard approach. The second source of uncertainty is a possible deviation of the ratio $(N_\nu + N_{\bar{\nu}})/N_\gamma$ from the canonic value 3/11. For example massive particles with annihilation or decay life-time in the interval $1 \div 10^4$ sec could increase N_γ without distortion of the spectrum of the microwave background radiation. The ratio N_ν/N_γ can be made in this way $2 \div 3$ times smaller without breaking successful results of the primordial nucleosynthesis theory.

Eq. (0.37) for the neutrino number density is obtained for ν with one polarization state only i.e for ν_L and $\bar{\nu}_R$. If $m_\nu \neq 0$ the particles can have both left-handed and right-handed polarization. It could seem that the result (0.37) should be divided by 2 to take into account the neutrinos with improper polarization. But this is not so because the light ν_R and $\bar{\nu}_L$ are not in thermal equilibrium with the primeval plasma at $T < 100$MeV and the entropy delivery to the neutrinos with proper polarization makes $N_{\nu_R} < N_{\nu_L}$.

The obtained above upper bound on m_ν is applicable with slight modification to any particles x which were relativistic at the moment of decoupling. If the interactions of these particles are weaker than interactions of neutrinos they are decoupled at higher temperature than ν and so with larger number of effective degrees of freedom in the primeval plasma. So their relative concentrations N_x/N_γ at the present time should be $K_x/K_\nu \approx K_x/10$ smaller than that of ν and correspondingly the bound on their mass is weaker by the same factor. Here K_x is the number of effective degrees of freedom in the primeval plasma at the moment of the decoupling of x-particles. If x-particles were in thermal equilibrium when the Universe reheated after the the end of inflation and decoupled soon after that their mass should not be greater than a few KeV because K_x is probably not larger than 10^3. If however the interactions of x are so weak that they are not abundantly produced after inflation is over the limit on their mass depends upon their production and can be considerably weaker.

Let us consider now the particles which were nonrelativistic at the decoupling. If the products of annihilation of $x\bar{x}$ are in thermal equilibrium and the distribution of x in energy has also the equilibrium form integro-differential eq. (0.22) describing the evolution of the distribution function $n_x(p)$ can be reduced to the ordinary differential equation for the particle number density $N_x = \int d^3p n_x(p)/(2\pi)^3$. Indeed let pairs of $x\bar{x}$-particles are produced and annihilated in reactions of the type $x + \bar{x} \to f$ where f is generically multiparticle and, what is essential, equilibrium state. The collision integral in this case has the form

$$S_x = -\frac{(2\pi)^4}{2E}\sum_f \int \frac{d^3p_{\bar{x}}}{(2\pi)^3 2E} d\nu_f \delta^4(p_x + p_{\bar{x}} - p_f) \mid A(x + \bar{x} \to f) \mid^2 \left(n_x^2 - \prod_f n_f\right) + S_{el} \tag{0.41}$$

(see eq. (0.24)). Here S_{el} describes the processes of elastic scattering. It was assumed that $n_x = n_{\bar{x}}$ and that T-invariance is not broken so that $\mid A(x + \bar{x} \to f) \mid^2 = \mid A(f \to x + \bar{x}) \mid^2$ (the particle velocities should be reversed but this is not essential). The role of the Fermi/Bose corrections has been analyzed by Dolgov and Kainulainen (1993) and it was shown that depending upon the ratio T_f/m, their magnitude varies from about 10% to negligibly small.

Since the state f is equilibrium then in the limit of the Boltzman statistics

$$\prod_f n = \exp\left(-\frac{E_x + E_{\bar{x}}}{T}\right) = n_x^{eq} n_{\bar{x}}^{eq} \tag{0.42}$$

where n_x^{eq} are the equilibrium distribution function of particles x. Note that $\prod_f n_f$ does not depend on the energies of the separate particles which form the state f. The distribution function $n_x(p_x, t)$ starts to deviate from the equilibrium one when the annihilation rate $\dot{N}_x/N_x \approx \sigma N_x$ becomes smaller than the expansion rate $\dot{a}/a = H$. The distribution in momentum however can maintain the equilibrium form because it is established by faster processes of elastic scattering with the rate $\sigma_{el} N_0$ where N_0 is the number density of light particles which are scattered on x and $\bar{x}$. We see below that the equilibrium with respect to annihilation is broken when $T/m_x < 1$ and thus $N_x \ll N_0$. So we can write

$$n_x = n_{\bar{x}} = \exp[(\mu - E)/T] \tag{0.43}$$

The nonequilibrium manifests itself in equality $\mu_x = \mu_{\bar{x}}$ whereas in equilibrium $\mu_x = -\mu_{\bar{x}} = 0$. The possible effects of charge symmetry breaking which could result

in $\mu_x \neq \mu_{\bar{x}}$ are neglected here but they are very essential in the kinetics of the generation of the baryon asymmetry.

If now we integrate the collision integral S_x (0.41) over $d^3p/(2\pi)^3$ the contribution of S_x^{el} disappear because of the kinetic equilibrium. Since the particles x are by assumption nonrelativistic (though this is not essential),

$$\int \frac{d^3p_x n_x}{(2\pi)^3 2E_x} = \frac{N_x}{2m} \tag{0.44}$$

Here N_x is the total number density with the account of the spin states. The integral over $d\nu_f$ gives by definition

$$\sum (2\pi)^4 \int d\nu_f \delta^4(p_x + p_{\bar{x}} - p_f) \mid A(x + \bar{x} \to f) \mid^2 = 4m_x^2 \sigma v \tag{0.45}$$

where v is the relative velocity (in the nonrelativistic limit) of the colliding x and $\bar{x}$ and σ is the total cross-section of $x\bar{x}$-annihilation averaged over the colliding particle spins. (We have assumed that the number density of x and $\bar{x}$ do not depend on the polarization.) If the annihilation proceeds in S-wave the product σv does not depend on the collision velocity. In the general case

$$\sigma v = (\sigma v)_0 v^{2l} = (\sigma v)_0 \frac{(p_x + p_{\bar{x}})^{2l}}{m_x^{2l}} \tag{0.46}$$

where l is the orbital momentum of the colliding particles.

Thus it follows from eq. (0.22) (see e.g. Zel'dovich and Novikov, 1975) that

$$\dot{N}_x + 3HN_x = -\langle \sigma v \rangle \left(N_x^2 - N_{x,eq}^2 \right) \tag{0.47}$$

where $N_{x,eq}$ is the equilibrium number density of particles x (0.9,0.7) and $\langle \sigma v \rangle$ is the the temperature average of σv:

$$\langle \sigma v \rangle = N_x^2 \int \frac{d^3p_x}{(2\pi)^3} \frac{d^3p_{\bar{x}}}{(2\pi)^3} n_x n_{\bar{x}} \sigma v = C_l (\sigma v)_0 (T/m_x)^l \tag{0.48}$$

where

$$C_l = 2^l \int_0^\infty dx\, dy e^{-x^2-y^2} \int_{-1}^{+1} d\zeta \left(x^2+y^2+2xy\zeta\right)^l x^2y^2/(l=0) \qquad (0.49)$$

In particular $C_0 = 1$ and $C_1 = 6$.

Eq. (0.47) can be used for calculation of the limiting for $t \to \infty$ concentration of stable relics of big bang. The conditions of its validity has been formulated in the derivation. This equation can be solved numerically to determine the frozen number density of massive relic particles which survived after big bang but in many cases a very simple approximate expression is sufficient

$$\frac{N_x}{N_\gamma} \approx \frac{1}{\sigma_{ann} m_{Pl} m_x} \qquad (0.50)$$

This result permits to evaluate quickly the mass density of the relic particles.

We have considered the case when particles are burned in the two-body collisions $x + \bar{x} \to all$. It is possible that particles can be burnt only in three-body collisions. The physical example of such case has been considered by Okun (1980) who discussed the existence of new particles with large radius of confinement. The model predicts neutral massive particles θ which disappear in the course of the Universe expansion through the reaction $3\theta \to 2\theta$. The residual concentration of relic particles disappearing in three-body collisions is (Dolgov, 1980)

$$r_\theta = N_\theta/n_\gamma \approx 100(m_\theta/m_{Pl})^{1/2}(m_\theta/T_f)^2(\Gamma m_\theta^5)^{-1/2} \qquad (0.51)$$

where

$$\Gamma = (2m_\theta)^{-3} \int \mid A(3\theta \to 2\theta) \mid^2 d\tau_2 \qquad (0.52)$$

is the probability of the transition $3\theta \to 2\theta$ normalized to unit number density and T_f is the freezing temperature,

$$\frac{m_\theta}{T_f} \approx 21 + \frac{1}{2}\left[\ln(\Gamma m_\theta^5) - \ln\frac{m_N}{m_\theta} - \frac{1}{2}\ln K\right] \qquad (0.53)$$

Note that the asymptotic value of the concentration is proportional to m_{Pl}^{-1} for two-body burning and to $m_{Pl}^{-1/2}$ for three-body burning.

Now let us consider a few simple examples on using these results. We have shown that the number density of particles in the comoving volume tends to a constant value in the course of the universe expansion. This phenomenon is called the concentration freezing. The calculated number density of cosmic relics can be used for the derivation of the cosmological bounds on their masses or the interaction strength in the same way as it has been done for neutrinos.

If one tries to apply eq. (0.50) to nucleons (Zel'dovich, 1965; Chiu, 1965) the result would be discouraging

$$r_{Bf} = r_{\bar{B}f} \approx 10^{-19} \tag{0.54}$$

The cross-section of $N\bar{N}$-annihilation at small energies is taken to be $\sigma v \approx 10^{-25} cm^2$ and the freezing temperature (that is the temperature when the annihilation effectively stopped) is $T_1 \approx m_N/45 \approx 25\text{MeV}$.

Astronomical data give much larger baryonic number density, $N_B > 10^{-10} N_\gamma$. This is an extra evidence in favor of the baryonic excess in the Universe at least at sufficiently small temperatures $T < m_N$. The number density of antibaryons in the case of baryonic excess should be exponentially small.

The analogous considerations have been used by Zel'dovich, Okun, and Pikelner (1965) for the evaluation of the concentration of the relic quarks in nature if the latter could exist as free particles and their absence in experiment were explained by the large mass. Assuming that $\sigma(q\bar{q}) \approx \sigma(N\bar{N})(m_N/m_q)^2$ we get

$$r_{qf} \approx (m_q/m_N) r_{Bf} \approx 10^{-19}(m_q/m_N) \tag{0.55}$$

The baryon asymmetry does not considerably change the result because the extra quarks should disappear in the reaction $qq \to B\bar{q}$.

Eq. (0.55) shows that if there existed free quarks their concentration relative to that of nucleons would be as large as 10^{-10}. This huge value is excluded by experiments on the search of free quarks in nature which give $10-15$ orders of magnitude smaller upper bounds. It is a strong evidence in favor of quark confinement.

The cosmological bounds on heavy particle masses are sensitive to the dependence of the annihilation cross-section on the masses of the colliding particles.

If gauge theory is valid then in the energy interval higher than the masses of intermediate bosons the cross-section is

$$v\sigma_{ann} \approx \pi\alpha^2/m_x^2 \tag{0.56}$$

Here m_x is the mass of the annihilated particles. In this case cosmology gives an upper limit on m_x. We assume in what follows that the energy density in the Universe is equal to the closure density, $\rho = \rho_c$. In this case the bound on m_x follows from the condition $\rho_x < \rho_c$.

The bound turns upside-down for the particles with masses smaller than the masses of intermediate bosons. In this case the annihilation cross-section is proportional to m_x^2 and the condition $\rho < \rho_c$ leads to a lower bound on m_x. Possible particles of this type could be hypothetical stable heavy neutral leptons L. If they possess the same interaction as neutrino differing only in mass the cross-section of their annihilation is

$$\sigma(L\bar{L} \to all)v = CG_F^2 m_L^2/6\pi \tag{0.57}$$

where $G_F = 10^{-5}m_N^{-2}$ is the Fermi coupling constant and the constant factor C is determined by the number of the possible annihilation channels. For $\sin^2\theta_W = 0.25$ each charged lepton with $m < m_L$, each neutrino, each upper quark, and each lower quark contributes into C respectively 1, 2, 2/3, and 13/24. The quark contribution is multiplied by 3 to take into account three quark colors. Thus $C \approx 10$.

Repeating the previous considerations but with cross-section (0.57) we obtain (Lee and Weinberg, 1977; Vysotsky, Dolgov, and Zel'dovich, 1977):

$$\frac{m_L}{m_N} > 1.8h_{100}^{-1} \tag{0.58}$$

The result is obtained for $K = 205/4$ and $N_{\gamma 0}/N_{\gamma f} = 13$ which corresponds to three types of left-handed neutrinos, electrons, muons, photons, 8 gluons, and u- and d-quarks with their antiparticles in the primeval plasma at the moment of freezing at $T_f = m_L/20$. Thus stable neutrinos should be roughly speaking either lighter than 30 eV or heavier than 2 GeV. Now we know even without cosmology that there are no new neutrinos in the forbidden mass interval because LEP data permits only 3 relatively light neutrinos and all three of the known ones are lighter than 2 GeV.

This approach is applicable for any stable particle. In particular in this way one can obtain the bound on the mass of the lightest supersymmetric particle which is possibly stable. Unfortunately the result depends on the masses of the superpartners which determine the annihilation cross-section. The details can be found in review by Sarkar (1985) or in paper by Kane and Kani (1986).

If the annihilation proceeds due to long-range forces the perturbation theory cross-section should be multiplied by the Coulomb factor (Sakharov, 1948):

$$\sigma_0 \to \sigma_0 \frac{2\pi\alpha/v}{1 - exp(-2\pi\alpha/v)} \tag{0.59}$$

This correction is essential for $2\pi\alpha > v \approx (3T/m)^{1/2}$. For electromagnetic interactions with $\alpha = 1/137$ this condition is not valid at $T = T_f \approx m/\ln W$. However in the case of strong interactions e.g. in annihilation of heavy quarks this cross-section rise can be essential and one have to use the given above equations with $l = 1/2$.

The bounds which have been discussed in this section can be reformulated as the lower bound on the annihilation cross-section (Dolgov and Zel'dovich, 1980):

$$\sigma v \geq 3 \cdot 10^{-37} cm^2 h_{100}^{-2} (N_{\gamma f}/N_{\gamma 0}) K_f^{1/2} \tag{0.60}$$

The surprising feature of the cosmological result is that it is a lower bound while laboratory experiments give upper bounds on cross-sections.