Scalar cosmological perturbations and gauges

Jose M. Torres López

May 8th, 2021, University of Rochester

Abstract

The flat Friedman-Robertson-Walker universe is the simplest approximation to the expanding nature of our Cosmos, but it fails to account for the anisotropies and inhomogeneities existing in it. The departing from this smoothness is important only at the correction level, and is therefore incorporated in the theory through first-order perturbations to the metric. In this paper, the general form of such perturbations, characterized by only four independent scalar functions, will be introduced. Then, the transformation relations for these functions under a change of reference frame, also referred to as gauge transformations, will be derived.

Introduction

The study of a universe model begins with the specification of its metric. One example is the expanding metric developed by Friedman, Robertson, and Walker during the 1920s and 1930s, which can be obtained from the Minkowski analogue through scaling of spatial distances by the variable parameter $a(t)$, as illustrated by Fig. 1 below, taken from [1] Scott Dodelson's *Modern Cosmology* (2003), page 26. Clearly, the flatness property is inherited in this way by the new metric -as can be shown by a straightforward application of Einstein's applications to find the associated energy-stress tensor.

$$g_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & a^2(t) & 0 & 0 \\ 0 & 0 & a^2(t) & 0 \\ 0 & 0 & 0 & a^2(t) \end{pmatrix}$$

Fig. 1. Matrix representation of the FRW metric. From Dodelson [1].

Of course, the Universe that we inhabit today is far from smooth (at least up to the galactic scale), as illustrated by the extensions of nearly void space that separate planets, stellar systems, and so on, or by the fluctuations of the CMB radiation (even though, at larger scales, the relative irregularities become smaller and smaller). Following linear perturbation theory, our objective then is to introduce small scalar perturbations (from now on, simply "perturbations") that account for the aforementioned irregularities in the cosmic distribution of energy. This approach is depicted by [2] Fig. 2 below (Kurki-Suonio. 2020), and it is mathematically well-defined by [3]; for a Robertson-Walker universe, solutions of the linearized field equations may be interpreted as linearizations of corresponding solutions for the nonlinear equations.

I will show first that four independent and arbitrary functions suffice to describe all possible perturbations in a general way. Being scalar quantities, these are not affected by coordinate transformations; but the metric tensor is. To retain the same four-perturbation structure of the metric, we have to adjust for this change through an appropriate redefinition of all four scalar functions.

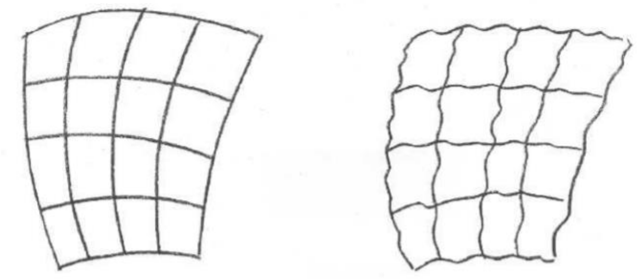

Fig. 2. Comparison between the flat and the perturbed spacetimes. From Kurki-Suonio [2].

In short, fixing a specific reference frame determines the value of the scalar functions and corresponds to a choice of gauge. Since different areas of Cosmology have distinct preferred gauges (because of the relative simplicity of resultant equations or complexity of related numeral calculations), these coordinate transformations become extremely useful tools as bridges between unequal gauges.

General form of first-order perturbations

We begin our derivation with the addition to the FRW metric of generic perturbations, but only those that are independent, i.e., which number cannot be reduced through the redefinition of a product or sum of more than one perturbations into a single one. For example, if C and D are scalar functions, their product CD can be combined into another generic function $N \equiv CD$, so C and D are not independent from this most efficient point of view. We will show that this consideration greatly limits the number of ways in which such quantities might enter the metric above. To do that, we follow the work of [4] Mukhanov, Feldman, and Brandenberger (1991), pp. 210-212.

For a start, the extension of the purely time component, $g_{00} = -1$, to general first-order perturbations requires the introduction of a single function A . Then, our expression becomes $g_{00} = -(1 + 2A)$. The correction term is $\delta g_{00} = -2A$, with the negative sign and factor of 2 being merely notational convention. Note that $A = A(x^0, \bar{x})$ is a function of time and space, so it shifts (slightly) the value of g_{00} possibly by a different amount at each point. The same applies to the remaining three functions, to be defined next, and their corresponding metric components.

In the exact same way, each of the terms $g_{0i} = 0$ (and g_{i0}, by symmetry of the metric tensor) can be most generally perturbed using one function. However, we note that these three terms can be expressed as the three covariant derivatives of a single scalar function B, which therefore suffices to accommodate all perturbations relevant to the spatial-time components g_{0i} (and g_{i0}). These covariant derivatives, denoted by $B_{|i}$, are taken along the i'th coordinate and over the hypersurface obtained by fixing a constant time with respect to the underlying FRW metric; they can also be interpreted as Lie derivatives, but in our case (derivative taken over flat background space, here FRW) they simplify into ordinary partial derivatives. The corresponding components are accordingly updated by $\delta g_{0i} = -B_{|i}$ to $g_{0i} = 0 + \delta g_{0i} = -B_{|i}$.

Having worked out in detail the first two cases above, we can directly identify the corresponding operations to be performed on the components of the type g_{ij}. Simply observe that scalar quantities might be added in two independent ways: by multiplying the non-zero diagonal components, as in g_{00} with A, or by adding the partial derivatives of a scalar function, as in g_{0i} with B. In this case, however, the derivatives will have to be of second order in order to accommodate the nine different components required for the nine g_{ij} definitions. Generality requires us to take both of these extensions into account, and we accordingly define the functions ψ, to be multiplied, and E, to be differentiated, which enter the metric as follows: $g_{ij} = a^2(\delta_{ij}[1+\psi] - 2E_{|ij})$. The rest of the paper deals with the resulting metric, summarized compactly by Eq. [1] below:

$$g_{00} = -(1+2A)$$

$$g_{0i} = -aB_{|i}$$

$$g_{ij} = a^2\left(\delta_{ij}[1+\psi] - 2E_{|ij}\right) \qquad [1]$$

Gauge transformations

In this section, we are interested in studying how the functions A, B, ψ, E just introduced must be adapted to accommodate the transformation of the metric under an arbitrary change of reference frame. Following chapter 5.5 of Dodelson [1], we consider an infinitesimal change of reference frame, which can be immediately exponentiated to generate the non-infinitesimal case. We observe that the most general expression of this type can be written in terms of the derivatives of a four-dimensional function $\mathcal{E}$, taken to be of the same order of magnitude as the scalar functions in the metric:

$$x^\mu \rightarrow \tilde{x}^\mu = x^\mu + \mathcal{E}_{|\mu}(x^0, \bar{x}) \qquad [2]$$

We can easily differentiate these equations to find how the transformed coordinates depend on the base coordinates:

$$\frac{\partial \tilde{x}^\mu}{\partial x^\nu} = \delta^\mu{}_\nu + \mathcal{E}_{|\mu\nu}$$

[3]

At this point, we need a relation between the original and the transformed metric tensors, $g_{\alpha\beta}$ and $\tilde{g}_{\alpha\beta}$, which we can find by remembering that the space-time distance given by the metric is an invariant quantity. In this way we get the equality:

$$\tilde{g}_{\alpha\beta}(\tilde{x})d\tilde{x}^{\alpha}d\tilde{x}^{\beta} = g_{\mu\nu}(x)dx^{\mu}dx^{\nu} \qquad [4]$$

Now, we can replace the differential elements of the transformed coordinates on the left-hand side, by means of the chain rule together with an application of Eq. [2]:

$$g_{\mu\nu}(x)dx^{\mu}dx^{\nu} = \tilde{g}_{\alpha\beta}(\tilde{x})\left(\frac{\partial\tilde{x}^{\alpha}}{\partial x^{\mu}}dx^{\mu}\right)\left(\frac{\partial\tilde{x}^{\beta}}{\partial x^{\nu}}dx^{\nu}\right)$$

$$= \tilde{g}_{\alpha\beta}(\tilde{x})\left[\left(\delta^{\alpha}{}_{\mu} + \varepsilon_{|\alpha\mu}\right)dx^{\mu}\right]\left[\left(\delta^{\beta}{}_{\nu} + \varepsilon_{|\beta\nu}\right)dx^{\nu}\right]$$

$$\rightarrow \quad g_{\mu\nu}(x) = \tilde{g}_{\alpha\beta}(\tilde{x})\left(\delta^{\alpha}{}_{\mu} + \varepsilon_{|\alpha\mu}\right)\left(\delta^{\beta}{}_{\nu} + \varepsilon_{|\beta\nu}\right) \qquad [5]$$

Distributing all the terms and then neglecting the last one for being of second order:

$$g_{\mu\nu}(x) = \tilde{g}_{\mu\nu}(\tilde{x}) + \tilde{g}_{\mu\beta}(\tilde{x})\,\varepsilon_{|\beta\nu} + \tilde{g}_{\alpha\nu}(\tilde{x})\,\varepsilon_{|\alpha\mu} + \tilde{g}_{\alpha\beta}(\tilde{x})\,\varepsilon_{|\alpha\mu}\,\varepsilon_{|\beta\nu}$$

$$= \tilde{g}_{\mu\nu}(\tilde{x}) + \tilde{g}_{\mu\beta}(\tilde{x})\,\varepsilon_{|\beta\nu} + \tilde{g}_{\alpha\nu}(\tilde{x})\,\varepsilon_{|\alpha\mu} \qquad [6]$$

The above formula relates in a simple way $g_{\mu\nu}(x)$ and $\tilde{g}_{\mu\nu}(\tilde{x})$. However, note that the argument of these two tensors are different. In order to compare the tensors, we want to express them as functions of a common set of coordinates, say x. With this in mind, we use a Fourier expansion on $\tilde{g}_{\mu\nu}(\tilde{x})$ around x and retain only the first two terms (the rest have order greater than one):

$$\tilde{g}_{\mu\nu}(\tilde{x}) = \tilde{g}_{\mu\nu}(x) + \frac{\partial\tilde{g}_{\mu\nu}}{\partial x^{\gamma}}(\tilde{x}^{\gamma} - x^{\gamma})$$

$$= \tilde{g}_{\mu\nu}(x) + \varepsilon_{|\gamma}\frac{\partial\tilde{g}_{\mu\nu}}{\partial x^{\gamma}} \qquad [7]$$

Substituting this back into Eq. [6], we get the final transformation law to be evaluated:

$$g_{\mu\nu}(x) = \tilde{g}_{\mu\nu}(x) + \tilde{g}_{\mu\beta}(\tilde{x})\,\varepsilon_{|\beta\nu} + \tilde{g}_{\alpha\nu}(\tilde{x})\,\varepsilon_{|\alpha\mu} + \varepsilon_{|\gamma}\frac{\partial\tilde{g}_{\mu\nu}}{\partial x^{\gamma}} \qquad [8]$$

Recall that the definitions of our functions A, B, ψ, and E are incorporated in the above expression through the term $g_{\mu\nu}$, as shown in Eq. [1]. Similarly, we want to express the transformed metric $\tilde{g}_{\mu\nu}$ in the exact same form by using adequate redefinitions of the scalar functions: $\widetilde{\mathrm{A}}$, $\widetilde{\mathrm{B}}$, $\widetilde{\psi}$, and $\widetilde{\mathrm{E}}$. The different components of Eq. [8] can then be evaluated explicitly to relate the new scalars to the original ones, and this will be the topic of the remaining of the paper.

Transformation of A

For this first computation, we must focus on the $_{00}$ component of Eq. [8], so that A appears. The left-hand side of this equality becomes – $(1+2A)$. On the right-hand side, the first term is given, accordingly, by $-\left(1+2\tilde{A}\right)$. The second one is: $\tilde{g}_{0\beta}(\tilde{x})\,\mathcal{E}_{|\beta 0}=-\left(1+2\tilde{A}\right)\left(\mathcal{E}_{|00}\right)-aB_{|i}\,\mathcal{E}_{|i0}\;=-\mathcal{E}_{|00}$, where we ignore higher-order terms in the last step; moving forward, I will do this without further comment. The two remaining terms, from left to right, are given by $\tilde{g}_{\alpha 0}(\tilde{x})\,\mathcal{E}_{|\alpha 0}=-\left(1+2\tilde{A}\right)\left(\mathcal{E}_{|00}\right)-aB_{|i}\,\mathcal{E}_{|0i}=-\mathcal{E}_{|00}$ and $\mathcal{E}_{|\gamma}\frac{\partial\tilde{g}_{00}}{\partial x^{\gamma}}=-\mathcal{E}_{|\gamma}\frac{\partial(1+2\tilde{A})}{\partial x^{\gamma}}=-\mathcal{E}_{|\gamma}\frac{\partial(2\tilde{A})}{\partial x^{\gamma}}=0$. The final transformation relation follows:

$$–(1+2A)=\;-\left(1+2\tilde{A}\right)-\mathcal{E}_{|00}-\mathcal{E}_{|00}$$

$$A\rightarrow\;\tilde{A}=A-\mathcal{E}_{|00}\qquad\qquad [9]$$

Transformation of B

We now consider the $_{0\mathrm{i}}$ equation, whose left hand side is simply $g_{0i}\;=\;-aB_{|i}$. Evaluate the other side:

$$\tilde{g}_{0i}(x)+\tilde{g}_{0\beta}(\tilde{x})\,\mathcal{E}_{|\beta i}+\;\tilde{g}_{\alpha i}(\tilde{x})\,\mathcal{E}_{|\alpha 0}+\;\mathcal{E}_{|\gamma}\frac{\partial\tilde{g}_{0i}}{\partial x^{\gamma}}$$

$$= -a\widetilde{\mathrm{B}}_{|i} - \left[\left(1 + 2\tilde{A}\right)\mathcal{E}_{|0i} + a\widetilde{\mathrm{B}}_{|j}\, \mathcal{E}_{|ji}\right]$$

$$-\left[a\widetilde{\mathrm{B}}_{|i}\, \mathcal{E}_{|00} - a^2\left(\delta_{ji}[1 + \psi] - 2E_{|ji}\right)\mathcal{E}_{|j0}\,\right] + \mathcal{E}_{|\gamma} \frac{\partial(-\, a\widetilde{\mathrm{B}}_{|j})}{\partial x^\gamma}$$

$$= -a\widetilde{\mathrm{B}}_{|i} - \mathcal{E}_{|0i} + a^2\mathcal{E}_{|i0}$$

First, note that $\mathcal{E}$ is a smooth function, so that $\mathcal{E}_{|i0} = \mathcal{E}_{|0i}$, and make this substitution. Then, equate the resulting expression to $g_{0i} = -aB_{|i}$, and integrate both sides over the i'th coordinate. It is very important to notice that, in this case, knowledge of the three spatial derivatives of B (given by the three independent equations contained in the general expression above) is enough to characterize the transformation. The reason is that, by definition, only the derivatives of B are physically meaningful, so that the constant of integration is irrelevant and can be taken to be zero. A bit of simplification then leads to the following result:

$$B \rightarrow \tilde{B} = B + \mathcal{E}_{|0}\left[a - \frac{1}{a}\right] \qquad [11]$$

Transformation of E

There are two kinds of metric elements involving the spatial partial derivatives of the function E: the spatial diagonal entries and the spatial off-diagonal ones. For now, we can avoid the apparent coupling with the function ψ (which we relegate to the last section) by choosing the off-diagonal components ($i \neq j$), that are worked out as follows:

$$-2a^2E_{|ij} = \tilde{g}_{ij}(x) + \tilde{g}_{i\beta}(\tilde{x})\, \mathcal{E}_{|\beta j} + \tilde{g}_{\alpha j}(\tilde{x})\, \mathcal{E}_{|\alpha i} + \mathcal{E}_{|\gamma} \frac{\partial \tilde{g}_{ij}}{\partial x^\gamma}$$

[12]

$$= -2a^2\tilde{E}_{|ij} - \left[a\tilde{\mathrm{B}}_{|i}\,\mathcal{E}_{|0j} - a^2\left(\delta_{ik}[1+\tilde{\psi}] - 2\tilde{E}_{|ik}\right)\mathcal{E}_{|kj}\right]$$

$$- \left[a\tilde{\mathrm{B}}_{|j}\,\mathcal{E}_{|0i} - a^2\left(\delta_{kj}[1+\tilde{\psi}] - 2\tilde{E}_{|kj}\right)\mathcal{E}_{|ki}\right]$$

$$+ \mathcal{E}_{|\gamma}\frac{\partial}{\partial x^\gamma}\left[a^2\left(\delta_{ij}[1+\tilde{\psi}] - 2\tilde{E}_{|ij}\right)\right]$$

$$= -2a^2\tilde{E}_{|ij} - 2a^2\mathcal{E}_{|ij}$$

where we have again used the fact that $\mathcal{E}_{|ij} = \mathcal{E}_{|ji}$ to combine the two terms proportional to these factors. Consequently, we can integrate directly (sanity check: we have here 3x3 = 9 equations for the same number of second-order derivatives) to see that the scalar function E must be redefined as:

$$E \rightarrow \tilde{E} = E + \mathcal{E} \qquad [13]$$

Transformation of ψ

At this point, we have but one choice left: we have already studied all the components except those diagonal and spatial.

$$a^2\left([1+\psi] - 2E_{|ii}\right) = \tilde{g}_{ii}(x) + \tilde{g}_{i\beta}(\tilde{x})\,\mathcal{E}_{|\beta i} + \tilde{g}_{\alpha i}(\tilde{x})\,\mathcal{E}_{|\alpha i} + \mathcal{E}_{|\gamma}\frac{\partial\tilde{g}_{ii}}{\partial x^\gamma}$$

$$= a^2\left(1+\tilde{\psi} - 2\tilde{E}_{|ii}\right) - \left[a\tilde{\mathrm{B}}_{|i}\,\mathcal{E}_{|0i} - a^2\left(\delta_{ij}[1+\tilde{\psi}] - 2\tilde{E}_{|ij}\right)\mathcal{E}_{|ji}\right]$$

$$- \left[a\tilde{\mathrm{B}}_{|i}\,\mathcal{E}_{|0i} - a^2\left(\delta_{ji}[1+\tilde{\psi}] - 2\tilde{E}_{|ji}\right)\mathcal{E}_{|ji}\right]$$

$$+ \mathcal{E}_{|\gamma}\frac{\partial}{\partial x^\gamma}\left(a^2\left[1+\tilde{\psi} - 2\tilde{E}_{|ij}\right]\right)$$

$$= a^2\left(1+\tilde{\psi} - 2\tilde{E}_{|ij}\right) + 2a^2\mathcal{E}_{|ii} + \mathcal{E}_{|0}\left(2a\frac{\partial a}{\partial x^0}\right)$$

The term $2a^2E_{|ii}$ on the left-hand side together and $2a^2\mathcal{E}_{|ii}$ on the right-hand side cancel out with $-2a^2\tilde{E}_{|ij}$ if we apply Eq. [13] to the latter. Finally, we let $H \equiv \frac{1}{a}\,\frac{\partial a(x^0)}{\partial x^0}$

denote the Hubble expansion constant, and use it in the presentation of the last gauge transformation:

$$\psi \rightarrow \widetilde{\psi} = \psi - H\, \mathcal{E}_{|0} \qquad [14]$$

Conclusion and final remarks

This last section emphasizes the physical interpretation of the results obtained above. Besides, a glimpse of the general picture of linear perturbation and gauge theory will be provided. To begin with, we collect the laws derived in the last half of the paper to present them in a more compact form:

$$\begin{gathered} A \rightarrow \tilde{A} = A - \mathcal{E}_{|00} \\ B \rightarrow \tilde{B} = B + \mathcal{E}_{|0}\left[a - \frac{1}{a}\right] \\ E \rightarrow \tilde{E} = E + \mathcal{E} \\ \psi \rightarrow \widetilde{\psi} = \psi - H\, \mathcal{E}_{|0} \end{gathered} \qquad [15]$$

The analysis of these equations begins with the importance contributions of $\mathcal{E}$ and its derivatives. Indeed, $\mathcal{E}$ depends on the specific coordinate transformation, so it is straightforward to go from any current frame of reference to a different one that simplifies one or more of the gauge functions above. The clearest example is given by the E function: if it is different than zero, we may switch to another frame through a transformation characterized by $\mathcal{E} = -E$, so that $\tilde{E} = 0$ in the new system. But only the spatial derivatives of $\mathcal{E}$ are relevant in this case (for the third of the equations), because E itself is determined by $E_{|ij}$. Hence, the time derivative $\mathcal{E}_{|0}$ represents another independent degree of freedom that we can adjust to simplify the gauge. It is clear then that only $4 - 2 = 2$ independent gauge choices exist: the four coming from A, B, E, and

ψ; the two, from $\mathcal{E}, \mathcal{E}_{|0}$. In fact, it is possible to combine directly our four functions in order to construct pairs of gauge-independent values − see [4] Bardeen (1980).

Next, let's consider the reason why we could ignore vector and tensor perturbations in this paper, and still get sensible laws for their scalar counterparts. In short, the decomposition theorem states that these three different categories of perturbations evolve independently of each other, so that the results found would have been in no way affected had we added, for instance, a tensor perturbation to the metric in Eq. [1].

Finally, address the essential question about these equations. How do the scalar functions represent anisotropies and inhomogeneities, exactly? A quick glance to Eq. [1] provides the answer: they are present in different components of the metric, which therefore, for most combinations of values for the scalar functions will have different expressions in different directions. Consequently, distances will be contracted or expanded in a non-uniform way. In turn, this means that there exist irregularities in the underlying energy distribution, which curves space-time (giving rise to the metric) in the first place. In other words, our scalar functions are the mathematical representation of physical perturbations, such as the irregularities in the CMB radiation visible in Fig. 3.

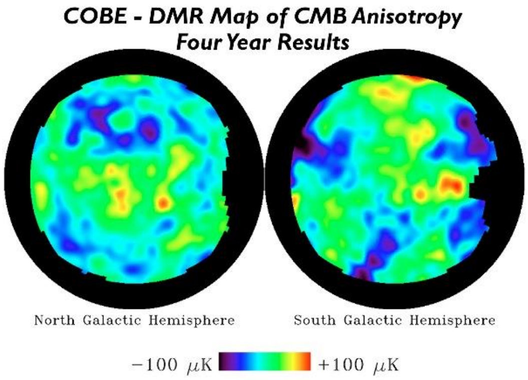
COBE - DMR Map of CMB Anisotropy
Four Year Results
North Galactic Hemisphere
South Galactic Hemisphere
−100 μK
+100 μK

Fig. 3. CMB anisotropies measured by NASA's COBE. From [5].

References

[1] Dodelson, S. *Modern Cosmology*. Academic Press. (2003).

[2] Kurki-Suonio, H. *Cosmological Perturbation Theory, part 1*. (2020).

[3] D'Eath. P. Ann. Phys. 98 (1976) 237.

[4] Mukhanov, V.; Feldman, H.; and Brandenberger, R. *Theory of cosmological perturbation,* pp. 210-212*. Physics Reports.* (1992).

[5] NASA / COBE Science Team. COBE Slide Set – High-Resolution Images.