AN ANALYTIC MODEL FOR THE RIEMANNIAN SPACE OF COLORS

By B. Lukács

Central Research Institute for Physics, Hungarian Academy of Sciences, Budapest*

(*Received April 18, 1986*)

A scale-invariant generalization of Weinberg's theory of the color space is given. A minimal analytic model is constructed, with Gaussian protomeric basis; the metric tensor possesses four independent Killing vectors with an U(1)⊗SO(1,1) symmetry group. The formalism is applied to dichromatic vision too, and a special model is shown, in which the color space is two-dimensional, however the dominant hues can be identified.

PACS numbers: 89.90.+n

1. *Introduction*

It is well-known that the (subjective) space of the colors is three-dimensional, but not Euclidean, if the distance of two colors is defined by the number of distinguishable colors between them [1]. (Such a definition is natural and usual, an example for similar distance definition in another context is Wootters' statistical distance [2, 3]). While the metric can be directly measured, it would be much more simple if the form of the line element were known and only some parameters should be measured. Some semi-empirical line elements have been, indeed, suggested (cf. e.g. Refs [1] and [4],) perhaps the most promising of them is Weinberg's proposition [5], establishing a close connection between the color metric and the natural selection. His startpoint is that the primary evolutionary gain of the color vision is that the individual can identify desirable or dangerous reflectants at various natural illuminations, which is an ability profitable for survival, indeed. Thus a theory of color vision must simply explain the relative constancy of the percepted color. However, there is a serious problem.

The experiments show that the color coordinates are additive for addition of lights, thus the color coordinates have to be formed from the incident intensity distribution $V(\lambda)$ as

$$v_i = \int_0^\infty A_i(\lambda)V(\lambda)d\lambda, \tag{1.1}$$

$$i = 0, 1, 2.$$

* Address: Central Research Institute for Physics, Hungarian Academy of Sciences, P.O.B. 49, 1525 Budapest, Hungary.

The simplest model explaining the form (1.1) is the three receptor model, in which (up to linear combinations of constant coefficients) $A_i(\lambda)$ are the sensitivities of the receptors. Now, the quantity, which should be observed for identifying the object is the reflectance function $R(\lambda)$, while the quantities, which can be observed, are the spectral distributions of the illuminating and reflected lights, $I(\lambda)$ and $V(\lambda)$, respectively.

Now, for color constancy, some coordinates belonging to $R(\lambda)$ should be formed from the coordinates belonging to $I(\lambda)$ and $V(\lambda)$, but there is no direct way to do this.

In order to resolve the apparent contradiction between the necessary and possible, Weinberg assumes that the brain approximates the incident spectral distributions by three parameter functions

$$V(\lambda) = e^{p^r P_r(\lambda)} \tag{1.2}$$

(henceforth we use the Einstein convention: there is a summation for indices occurring twice, above and below), with three fixed basic functions $P_i(\lambda)$. The functions of form (1.2) are called protomers. Now, three parameters belonging to $V(\lambda)$ can indeed be measured, namely the color coordinates defined by Eq. (1.1). So, if the brain wants to find a protomer to an arbitrary $V(\lambda)$, it can do it.

If the basis $P_i(\lambda)$ were complete for natural lights, i.e. all the lights occurring in the nature possessed spectral distributions of form (1.2), then the color constancy would be automatical, since the coefficients p^r are additive for multiplications of protomers. Thus, observing $I(\lambda)$ and $V(\lambda)$, the coefficients belonging to $R(\lambda)$ could be formed from the other two sets of coefficients.

While obviously Eq. (1.2) does not hold exactly for all the natural lights, it is possible to find the best basis. Then the color constancy is not absolute, but can be quite good, as far as the lights are not too far from the form (1.2) assumed by the brain when evaluating the incident lights.

Weinberg has shown that (if the basis is already fixed somehow), the color constancy is optimal if there is a proportionality between the sensitivity functions $A_i(\lambda)$ and the protomeric basis $P_i(\lambda)$:

$$A_i(\lambda) = C(\lambda)P_i(\lambda), \tag{1.3}$$

where $C(\lambda)$ is an arbitrary function. If Eq. (1.3) holds (and natural selection sees that it holds), then, in Weinberg's model, the metric of the protomeric space has a simple form:

$$ds^2 = \frac{1}{H}\frac{\partial^2 H}{\partial p^i \partial p^k} dp^i dp^k,$$

$$H = \int_0^\infty C(\lambda) e^{p^r P_r(\lambda)} d\lambda. \tag{1.4}$$

This gives the metric in the color space in an implicit way, since, from Eqs (1.1) and (1.3),

$$v_i = \frac{\partial H}{\partial p^i}. \tag{1.5}$$

Thus Weinberg's theory generates the metric by means of four functions of one variable, which is a big enough simplification compared to the original problem, namely to six functions of three variables, without a theory. If one knew $C(\lambda)$ (the luminous efficiency), and $P_i(\lambda)$ (the protomeric basis), then the distance of any two protomers could be calculated in a direct way, and that of any two colors could be calculated in an indirect way.

However, the real situation is not this, we are confronted with the inverse problem, how to get $C(\lambda)$ and $P_i(\lambda)$ from some measured distances, and it seems that to this goal there is no direct way, similarly to the quantum mechanical problem when the scattering potential should be calculated from the cross section. Even for a start, some guesses would be necessary for $C(\lambda)$ and $P_i(\lambda)$.

In fact, Weinberg's paper contains some such guesses, at least for the P_i's. Namely, he notes that the conic shape of the spectral loop suggests Gaussian protomers, i.e. the basis

$$P_0 = 1, \quad P_1 = \lambda, \quad P_2 = \lambda^2, \tag{1.6}$$

which is in accordance with well-founded expectations, Gaussian distributions being generally useful for overall representations of reasonable distribution functions. For $C(\lambda)$ he is definitely less optimistic [5].

Nevertheless, there is a problem. The conic shape of the spectral loop is consistent with Eq. (1.6), but does not lead to it. In the best case it leads to something similar, with an undefined function of λ instead of λ'. Now, this means that by transforming the wavelength scale one could arrive at a Gaussian basis, but then the behaviour of the fundamental quantities of the model in such a transformation should be studied too. Furthermore, there is a possibility that the spectral distributions of the lights in the nature would suggest something for $C(\lambda)$. However, a suggestion can be obtained only on the physical wavelength scale, not on an undefined one. Thus it may be profitable to remain at the physical scale, where even (1.6) is not necessary.

Since the situation is not quite clear, we want to get some orientation first. So in this paper we are going to do only three steps: to see the behaviour of the functions of the model with respect to a scale change; to get some suggestions for the functions from physical facts; and to construct such particular models when the metric can be analytically expressed, because in such cases one can see in this Riemann space.

In Section 2 we discuss the behaviour of the quantities generating the metric when performing a transformation of the physical variable ("wavelength"). The result is that the theory can be used on any physical scale if a further function of λ (called principal illumination here) is introduced; such a function guarantees the proper transformation laws, and, in fact, possesses some direct meaning. Sections 3–6 contain some arguments on the approximate forms of the wavelength functions of the theory for the realistic case, together with a discussion of the role of a free constant conformal factor of the formalism. In Section 7, using some simplifying assumptions discussed in the previous Sections, a minimal model is constructed, in which the metric tensor can be analytically expressed. Section 8 gives a brief analysis of the possible types of anomalous color vision in the framework of Weinberg's theory.

2. The scale invariance

Eqs (1.3)–(1.5) show that in Weinberg's theory (henceforth WT) the observables and their relations are generated by functions of one variable, the wavelength λ. Nevertheless, this variable is dummy, it does not appear in the final results. In addition, it is sure that the brain does not have to use the notion of wavelength for evaluating the signals of the receptors of the eye, because color vision did exist before Huyghens. So, one may require that the formalism keep its form when using anything convenient instead of λ (e.g. the frequency c/λ). Nevertheless, this requirement imposes some constraints on the formalism, since there must exist some invariant quantities in the theory.

Consider a group of illuminations with intensity distributions of form (1.2), and reflectants of similar reflectivities. Then the reflected lights possess the same forms too; all the incident lights will be protomers in this ideal system. Then the color constancy will be absolute. But in this way one can see that the fact that an intensity distribution is a protomer has some physical meaning. Now, consider a light which is a protomer, so for which Eq. (1.2) holds. Then, introducing a new "wavelength" scale λ', $V(\lambda)$, being a distribution, changes. On the right hand side both p^i and $P_i(\lambda)$ may change, but the equation must hold again, because the light is a protomer. Thus

$$V'(\lambda') = V(\lambda)\,\frac{d\lambda}{d\lambda'} = e^{p^{r'}P_{r'}(\lambda')}. \tag{2.1}$$

By combining Eqs (1.2) and (2.1) one gets

$$\left(p^r P_r(\lambda) - \ln\frac{d\lambda'}{d\lambda}\right)_{\lambda=\lambda(\lambda')} = p^{r'}P_{r'}(\lambda') \tag{2.2}$$

which is a constraint for the transformation law of the basis.

Another quantity, which has to be independent of the wavelength scale used in the theory, is the color coordinate v_i, produced by the receptors. But then, from Eq. (1.1):

$$A_{i'}(\lambda') = A_i(\lambda(\lambda')). \tag{2.3}$$

Expressing A_i by means of Eq. (1.3) one obtains for $P_{i'}$:

$$P_{i'}(\lambda') = [C'(\lambda')]^{-1i}[C(\lambda)P_i(\lambda)]_{\lambda=\lambda(\lambda')}, \tag{2.4}$$

with a still undefined transformation law for $C(\lambda)$. Writing this into Eq. (2.2) the result is as follows:

$$(p^{r'}CC'^{-1} - p^r)P_r = -\ln\frac{d\lambda'}{d\lambda}. \tag{2.5}$$

But this relation cannot hold for arbitrary $\lambda'(\lambda)$ and for arbitrary value of p^i, as it can be seen by differentiating it with respect of p^i. So one cannot impose such transformation law on $P_i(\lambda)$ and $C(\lambda)$ that relation (1.2) be scale-invariant, although it should be such. Nevertheless, the whole formalism can easily be made scale-invariant in a natural way.

Let us modify the definition of protomers as follows:

$$V(\lambda) = W(\lambda)e^{p^T Pr(\lambda)} \tag{2.6}$$

by introducing a further fixed function W (let us postpone the physical meaning of $W(\lambda)$ for a moment). Then, by requiring that Eq. (2.6) hold after a scale transformation, and v_i remain the same, one gets the transformation laws

$$W' = W(d\lambda'/d\lambda)^{-1}, \quad P_{i'} = P_i, \quad p^{i'} = p^i, \quad C' = C. \tag{2.7}$$

That is, $W(\lambda)$ is an intensity distribution. Since these transformation laws are of natural form, one may accept Eq. (2.6) as the natural scale-invariant generalization of the definition of protomers.

Obviously, the protomers defined according to Eq. (2.6) do not form a group under multiplication or division, but they do not have to. Namely, consider the real process resulting in observation. There is an illuminating spectral intensity $I(\lambda)$, which the brain approximates by a protomer:

$$I(\lambda) \simeq W(\lambda)e^{p^T{}_I Pr(\lambda)}. \tag{2.8}$$

This illumination is reflected by an object possessing a reflectance coefficient $R(\lambda)$, so the incident light has the intensity distribution

$$V(\lambda) = R(\lambda)I(\lambda). \tag{2.9}$$

This distribution is approximated by a protomer too:

$$V(\lambda) \simeq W(\lambda)e^{p_V{}^T Pr(\lambda)}. \tag{2.10}$$

The input data are the color coordinates belonging to I and V, thus the brain can determine p_I^i and p_V^i. Hence it obtains the reflectance parameteis as

$$p_R^i = p_V^i - p_I^i. \tag{2.11}$$

For this it is not necessary to assume that $R(\lambda)$ (approximately) possesses the form (2.6) too, and, in fact, it cannot possess, since $R(\lambda) = V(\lambda)/I(\lambda)$ is not an intensity distribution; being a ratio of distributions, in a scale transformation it changes as

$$R'(\lambda') = R(\lambda(\lambda')). \tag{2.12}$$

Thus, Eq. (2.11) implies that $R(\lambda)$ is approximated according to Eq. (1.2). Since in real processes the only operations are of form

$$V = IR; \quad R = V/I; \quad I = V/R,$$

the forms (2.6) for I and V, and the form (1.2) for R are kept.

In WT $R(\lambda)$ does not appear explicitly when generating the metrics or obtaining the conditions for maximal color constancy. Thus the whole calculation can be repeated with the (2.6) form of protomers. Since the calculation is lengthy, but a simple repetition, here we give only the final result. Eqs (1.3–5) remain valid except for a change in the equation for H. Thus, the maximal color constancy leads to

$$A_i(\lambda) = C(\lambda)P_i(\lambda), \tag{2.13}$$

the metric has the form

$$ds^2 = \frac{1}{H}\frac{\partial^2 H}{\partial p^i \partial p^k} dp^i dp^k, \quad H = \int_0^\infty C(\lambda)W(\lambda)e^{p^r P_r(\lambda)} d\lambda, \tag{2.14}$$

and the color coordinates belonging to the protomeric coordinates p^i can be obtained as

$$v_i = \frac{\partial H}{\partial p^i} \,. \tag{2.15}$$

Obseive that the generalized, scale-invariant theory possesses a gauge-type symmetry

$$p^{i'} = p^i + {}_0p^i, \quad W'(\lambda) = W(\lambda)e^{-{}_0p^r P_r(\lambda)}, \tag{2.16}$$

with constant ${}_0p^i$-s.

Now, we are in the position to attribute some physical meaning to $W(\lambda)$. Assume that the observer is in such a situation that he is not able to observe the illuminating light. Elementary experience shows that in such a situation the observer assigns a color to the reflectant, so an illumination of form (2.6) is still assumed by the brain. The details may depend on the partial knowledge about the situation. But it is natural to believe that some "principal illumination" $I_0(\lambda)$ is assumed when nothing is known. Shift the origin of the coordinate system p^i to this point by the gauge freedom (2.16), and then

$$I_0(\lambda) = W(\lambda). \tag{2.17}$$

That is, $W(\lambda)$ is the principal illumination, the presupposition of the brain about the most natural illumination.

WT offers one more natural possibility for generalization. In it, the distance is proportional to some measure of discrepancy Δ (cf. Eq. (9.6) in Ref. [5]). This discrepancy measures the deviation of the actual intensity distribution from its protomer (cf. Section 7 in Ref. [5]). Now, Δ is not uniquely determined in WT, but if it is small, there remains a simple scale change

$$\Delta \to c^2\Delta. \tag{2.18}$$

While this scale change is almost trivial, it occurs in the line element as a constant conformal rescaling

$$ds^2 \to c^2 ds^2, \tag{2.19}$$

which has measurable consequences. Such an arbitrary constant conformal factor is familiar in theories using Riemannian space, e.g. in General Relativity a line element remains a solution of the vacuum Einstein equation if multiplied by a constant [6].

3. The protomeric basis and the shape of the spectral loop

We have seen that the scale-invariant generalization of WT contains five unknown functions of the wavelength, the protomeric basis functions $P_i(\lambda)$, the luminous efficiency $C(\lambda)$ and the principal illumination $W(\lambda)$. While it is possible to determine all these functions from distance measurements, obviously it would be desirable to get as much of these functions as possible directly. Weinberg emphasizes that some functions can be determined from the color coordinates of the points of the spectral loop. Let us discuss this possibility.

By combining Eqs (1.1) and (2.13) one gets that the color coordinates of a sharp line of unit intensity

$$V(\lambda) = \delta(\lambda - \lambda_0) \tag{3.1}$$

can be given as

$$v_i(\lambda_0) = C(\lambda_0)P_i(\lambda_0), \tag{3.2}$$

which means three equations for four functions. However, Weinberg has proposed the assumption that a neutral amplification of lights does exist [5]. Namely, elementary experiences seem to indicate that a simple amplification of the illumination

$$I(\lambda) \to qI(\lambda) \tag{3.3}$$

does not alter the percepted colors as far as q is not too small or great. If such a neutral amplification does exist, then there exist such constants Q^i that

$$Q^r P_r = 1. \tag{3.4}$$

But then one of the P_i's can be expressed by means of the other two, and three constants. (The free constants could be eliminated by means of linear combinations of constant coefficients of the basic functions, but such an operation would lead to linear transformations in the color coordinates, and all the color measurements are performed in a special coordinate system defined by the Commission Internationale d'Eclairage [7].)

While the existence of neutral amplifications is not an a priori fact, it is a natural enough assumption, so here we shall not discuss if it is correct, but accept Eq. (3.4). Then Eqs (3.2) and (3.4) yield four relations for four functions, i.e. if the color coordinates of the spectral lines of unit intensity are known then the functions $P_i(\lambda)$ and $C(\lambda)$ can be expressed algebraically, up to three unknown constants.

Now, Weinberg is not too optimistic about the measurability of the total luminance, telling that the errors are substantial [5], and suggests to go as far as possible without using $C(\lambda)$. Then the normalized coordinates $v_i/(v_0+v_1+v_2)$ are to be used. Consider the "red" and "green" coordinates of the CIE system, x and y, respectively. Then a combination of Eqs (3.2) and (3.4) yields

$$x = \frac{Q^0 P_1}{1+(Q^0-Q^1)P_1+(Q^0-Q^2)P_2},$$

$$y = \frac{Q^0 P_2}{1+(Q^0-Q^1)P_1+(Q^0-Q^2)P_2}.$$

Since for the points of the spectral loop $x(\lambda)$ and $y(\lambda)$ are known, $P_1(\lambda)$ and $P_2(\lambda)$ can be expressed, up to the constants Q^i.

The spectral loop possesses a surprisingly perfect conic shape, except for the violet end. Let us assume that this is not an accident, forget the violet end, and look for the consequences.

Eq. (3.5) expresses x and y by means of two independent functions. Let us introduce some linear combinations of constant coefficients instead of P_1 and P_2. Then

$$x = \frac{A+B\varphi+C\psi}{U+V\varphi+W\psi}, \quad y = \frac{D+E\varphi+F\psi}{U+V\varphi+W\psi}, \tag{3.6}$$

where the capitals stand for constants, and the new functions φ and ψ are defined only up to linear transformations

$$\varphi \to K\varphi+L\psi+R, \quad \psi \to M\varphi+N\psi+S. \tag{3.7}$$

Now, the ideal spectral loop is a hyperbola, so

$$\alpha x^2+2\beta xy+\gamma y^2+\delta x+\varepsilon y+\zeta = 0,$$
$$\alpha\gamma-\beta^2 < 0. \tag{3.8}$$

Substituting the forms (3.6) into Eq. (3.8), and separating the terms of different wavelength dependences, one gets six relations among the five ratios α/ζ, β/ζ, γ/ζ, δ/ζ and ε/ζ. So generally the spectral loop cannot be conic, this is possible either for special values of the constants, or for special foims of the functions. The first case would seem accidental, then the very regular form of the loop would not be connected with anything important, so this possibility may be neglected by using Occam's razor. In the second case φ and ψ have to fulfill a quadratic relation. Hence, by using the free transformations (3.7) one gets three possible cases:

$$\psi^2 = \pm\varphi^2+1, \quad \psi = \varphi^2. \tag{3.9}$$

The third case may lead to Gaussian protomers, for which Section 1 gave some supporting arguments (see also in Ref. [5]), the other two cases do not lead to anything simple, so one may again use Occam's razor. Henceforth we accept the suggestion

$$\psi = \varphi^2. \tag{3.10}$$

However, the basis is Gaussian only if φ is a linear function of the wavelength (or of anything else important, as e.g. the frequency; physical considerations suggest that the relevant variable is either λ or $1/\lambda$).

With the constraint (3.10) the (x, y) coordinates given by Eq. (3.6) always lie on a cone, but it is a hyperbola only if

a) the denominator is linear in φ; or if

b) the denominator possesses two real roots.

Since we are looking for the linearity of φ in λ, transformations of form

$$\varphi \to a\varphi+b \tag{3.11}$$

are permitted, by means of them the canonical forms of these two cases can be written as

$$x = \frac{A}{\varphi} + B + C\varphi, \quad y = \frac{D}{\varphi} + E + F\varphi, \tag{3.12}$$

or

$$x = C + \frac{A+C+B\varphi}{\varphi^2-1}, \quad y = F + \frac{D+F+E\varphi}{\varphi^2-1}. \tag{3.13}$$

In the first case $\varphi(\lambda)$ can be expressed as

$$\varphi(\lambda) = C^{-1}\frac{m_1 \cdot \xi - \eta}{m_1 - m_2},$$

$$\xi = x(\lambda) - x_0, \quad \eta = y(\lambda) - y_0, \tag{3.14}$$

where m_i stand for the slopes of the two asymptotes, while x_0 and y_0 denote the coordinates of the crossing point of them, C can be made 1 by the transformation (3.11). In the second case

$$\varphi = -1 - \frac{K}{(y-y_0) - m_1(x-x_0) - K/2},$$

$$K = \frac{B(D+F)-(A+C)E}{A+B+C}. \tag{3.15}$$

Now, as far as the loop is a hyperbola (and it is quite hyperbolic, cf. Ref. [5]), both Eq. (3.12) and Eq. (3.13) can reproduce the shape, with the proper parameters. However, the Gaussian basis would mean that φ is linear in λ. By performing a fitting procedure one may determine the best linear formula for $\varphi(\lambda)$, which can be compared to the measured x, y values in order to decide the degree of Gaussian behaviour.

Observe that, by assuming the Gaussian nature of the basis as a tendency, one can definitely make a choice between the cases of linear and quadratic denominators. The shape, orientation and location of the hyperbola determine five of the constants A, B, C, D, E and F. In case (3.12) the sixth one is a trivial scale factor in φ, in the other case it appears in φ, and possesses a value when the linearity is the best. By compaiing this φ with that given by Eq. (3.14), and accepting the better, a choice has been done between Eqs (3.12) and (3.13), and all the constants have been determined.

At this point a detailed fitting procedure is not needed. Fig. 1 shows the color coordinates of the spectral loop for a particular linear $\varphi(\lambda)$ function, using the first case, compared to the tiue coordinates [8]. The tiue values are fairly reproduced between 510 mμ (bluish green) and 600 mμ (almost pure red). With quadratic denominator the global linearity seems worse.

When the denominator is linear, there is an asymmetry between the violet and red ends of the loop. In the same time, for the real loop, the violet end (where the Gaussian model

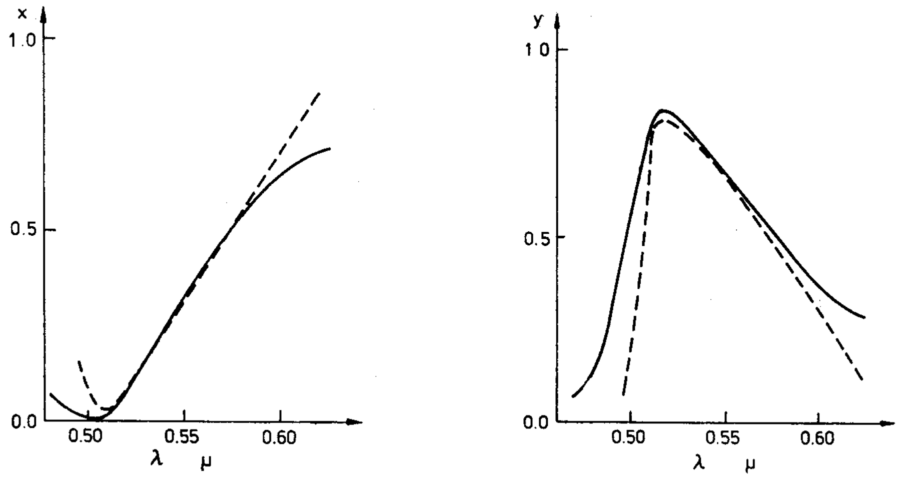

Fig. 1a

Fig. 1b

Fig. 1. The normalized color coordinates x (left) and y (right) of the spectral loop versus wavelength. The continuous lines represent the measured values [8], while the broken lines are valid for a Gaussian protomeric basis, according to Eq. (3.12). The constants $A \ldots F$ are so chosen that the place, shape and orientation of the spectral loop be the best possible, and the actual form of the linear function $\varphi(\lambda)$ is $\varphi = 7.803\lambda - 3.873$, where λ is measured in microns

tends to give a singularity) in fact differs from the red one (there is a turning back in violet). It is not clear, which is the reason and which is the consequence, if there is any correlation at all.

Obviously, φ cannot be linear for extreme values of λ. Such a function would lead to changing color coordinates even in the asymptotic region, but it would give very slight advantage compared to the problems of producing correct receptor sensitivities there. We will discuss the problem of the ends of the spectral loop later.

As for the direct expression of $P_i(\lambda)$, Eq. (3.5) can be inverted as

$$P_1 = \frac{x}{Q^0 + (Q^1 - Q^0)x + (Q^2 - Q^0)y}, \quad P_2 = \frac{y}{x} P_1. \tag{3.16}$$

Using Eq. (3.12) one gets that a linear function for $\varphi(\lambda)$ leads to Gaussian protomers if

$$Q^0 + E(Q^2 - Q^0) + B(Q^1 - Q^0) = 0,$$

$$F(Q^2 - Q^0) + C(Q^1 - Q^0) = 0. \tag{3.17}$$

Hence the ratios Q^1/Q^0, Q^2/Q^0, compatible with the (approximate) Gaussian nature of the real protomeric basis can be determined.

One can conclude that there exists such protomeric basis which reproduces the color coordinates of the spectral loop, and in the same time is Gaussian at least approximately, for the "important" region of the spectrum. It is a completely different question if this is the basis of the real color vision or not, i.e. if this basis is compatible with the measured distances in the color space, or not. Nevertheless, it would be very strange if the true basis were too far from the Gaussian one, and so the natural selection did not utilize this evident possibility.

4. The luminous efficiency

The luminous efficiency, $C(\lambda)$, plays a central enough role in WT (cf. Eqs (1.3–4)), and the situation remains similar in the scale-invariant generalization. Weinberg states that there is an intimate connection between $C(\lambda)$ and the luminance H, $C(\lambda)$ being the luminance of spectral lines of unit intensity; thus the luminance can be measured as some component of the color vector v_i iff 1 is a protomer.

Now, in the generalized theory these statements have to be reconsidered. First, H is defined for protomers only, and is a function of p^i's. Thus the luminance of a spectral line can be defined only if the spectral line is a protomer, which may or may not be true. Second, the value of H may be the same as that of C, but the sufficient condition will be slightly different. Namely, let us impose the assumption (3.4) on the particular form of the theory. Then one gets

$$H(p^i) = \int C(\lambda)W(\lambda)e^{p^r P_r(\lambda)}d\lambda = \int Q^r P_r(\lambda)C(\lambda)W(\lambda)e^{p^r P_r(\lambda)}d\lambda$$
$$= \int Q^r A_r(\lambda)W(\lambda)e^{p^r P_r(\lambda)}d\lambda = Q^r v_r(p^i), \tag{4.1}$$

where we have used Eqs (1.1) and (2.13). But v_i is a color coordinate, so it can be calculated for any intensity distribution which is a metamer of the chosen protomer. So, $C(\lambda)$ possesses the same value as $H(p^i)$ for the protomer of the spectral line of wavelength λ and of unit intensity. Similarly, the luminance can be somehow measured by means of color matching if the basic functions $P_i(\lambda)$ are such functions that Eq. (3.4) can hold for some constants Q^i. This does not mean that 1 is a protomer, because the principal illuminance may differ from unity (cf. Eq. (3.6)).

In Sect. 3 we saw that the assumption of a (nearly) Gaussian basis imposes some consistency conditions on the Q^i's. Thus, by this hypothesis, one may be able to determine the specific combination of the measurable color coordinates which is equal to the luminance, which is a possibility for measuring $C(\lambda)$.

If this measurement is technically difficult (as it is suggested by the pessimistic comments in Sect. 8 of Ref. [5]), then one needs a guess for $C(\lambda)$. Now, one may think that the luminance is in close relation with the total visual intensity of the light, determined by the relative sensitivity of the human eye. The latter one is measured [9, 10], it is roughly Gaussian for "visible" wavelengths, and decreasing exponential for near infrared. In order to make this more transparent, in Fig. 2 a difference of logarithms

$$\ln \frac{C(\lambda+\Delta\lambda)}{C(\lambda)}$$

is displayed, with $\Delta\lambda = 10$ mμ; it is decreasing linear function if C is Gaussian, and constant if C is exponential. One may conclude that $C(\lambda)$ is roughly (or "globally") Gaussian

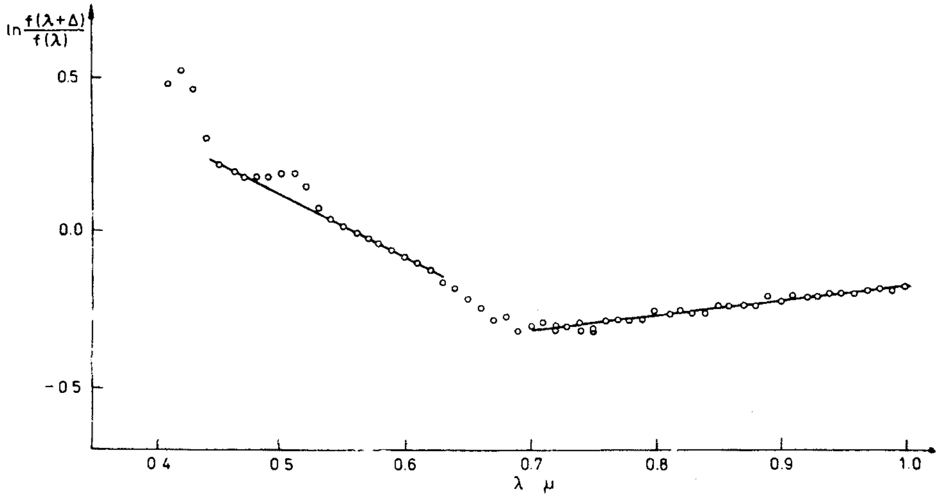

Fig. 2. The relative change of the sensitivity of the eye $f(\lambda)$ [9, 10] in the "visible" and near infrared regions. The displayed quantity is $\ln[f(\lambda+\Delta)/f(\lambda)]$, with $\Delta = 10$ mμ. If f is Gaussian, this curve is linear. More explanations in the text

between extreme blue and extreme red, with

$$\lambda_0 = 555 \text{ m}\mu, \qquad \sigma = 47 \text{ m}\mu \tag{4.2}$$

(although with serious deviation between 475 and 525 mμ). Nevertheless, these parameters should not be taken too seriously; there is some indication that they differ for different populations [11].

One can observe that nothing drastic happens at the red border of the color vision; the decrease of the sensitivity is gradual in the near infrared, and can be even slower than somewhere in the red. On the other hand, Ref. [10] states that in the infrared the color of the sufficiently intensive lights fluctuated between red and yellow. So it seems that there the protomeric basis is already far from Gaussian.

5. *The principal illuminance*

As we have seen, the scale-invariant formalism contains a new function $W(\lambda)$ occurring in all the protomers as a multiplicative factor. In Sect. 2 a suggestion was made that this function represents some preconception about the illumination, when it cannot be observed. Nevertheless, it is necessary to note that we have not proven that W plays this role indeed; in the form of the protomers $W(\lambda)$ appears from merely formal reasons. The formalism itself does not impose any constraint on $W(\lambda)$, e.g. it is even possible that $W(\lambda)$ is constant on the physical wavelength scale, although it would be difficult to find serious arguments for such a hypothesis.

Nevertheless, the formalism definitely needs $W(\lambda)$, and, if a formalism is proper for a problem (which is, of course, an open question for the present case), then its necessary components do play some role in the described mechanism. So, let us try to find some role for $W(\lambda)$.

In a natural environment the oveiwhelming majority of illuminations is a direct or indirect solar radiation. So, in order to get a satisfactory color constancy, the protomers have to be able to simulate the intensity distribution coming from the Sun in various hours of the day and under different meteorological circumstances. On the other hand, Gaussian protomers would be useful for simulating reflectances, and such protomers are slightly suggested by the conic shape of the spectral loop [5].

Now, consider the solar radiation. In first approximation it is a blackbody spectrum, with an intensity distribution

$$I(\lambda) = \frac{4\pi \hbar c^2}{\lambda^5} (e^{2\pi \hbar c/kT\lambda} - 1)^{-1}, \tag{5.1}$$

where $k = 1.38 \cdot 10^{-16}$ erg/grad, and T is the surface temperature, for the Sun, $T = 5770$ K [12]. Although this is a smooth curve of a single peak, it is far from a Gaussian. The peak is at cca. 500 mμ, with a half width 360 mμ, but even the peak is strongly asymmetric. In second approximation, there is a deviation from the blackbody spectrum, the violet end of the "visible" region is underpopulated in the solar radiation. The decrease becomes very steep at 390 mμ. Finally, there is a substantial atmospheric absorption. Between 290 and 760 mμ the absorption is moderate and a smooth function of the wavelength, but below 290 mμ it is practically complete by O_3 molecules, while in the near infrared the incident spectrum shows deep dentations caused by O_2 and H_2O [12]. The end of the red region coincides with a narrow and very deep absorption line of O_2.

Thus one can conclude that the sunlight, either direct or indirect, cannot be well approximated by Gaussian protomers, so such protomers cannot be expected from natural selection. However, observe that in the scale-invariant formalism protomers belong to lights, as illuminations and reflected spectra, and they possess forms according to Eq. (2.6), while reflectances are simulated by ratios of protomers, i.e. by functions given in Eq. (1.2). Then a Gaussian basis leads to Gaussian approximation for the reflectances, but there is a free function $W(\lambda)$ in the approximation of the illuminations. This function can be chosen in such a way that protomers be good approximations for the sunlight at least under the most important circumstances, e.g. at noon. The best agreement can be obtained if $W(\lambda)$ is just this particular distribution of the incident solar light, when one arrives again to Eq. (2.17), and then $W(\lambda)$ is, in fact, the intensity distribution of the principal illumination.

Nevertheless, the multiplicative function $W(\lambda)$ cannot be sufficient for simulating all the natural illuminations with a form (2.6). Namely, e.g. the relative strength of the absorption lines may change with the atmospheric situations and the distance of the Sun from the horizon (because of the change of the atmospheric mass along the path of the light). Now, one cannot expect that a changing pattern of narrow valleys could be simulated by the way (2.6), with smooth functions $P_i(\lambda)$, whether Gaussian or not. Thus the constant deviation from a Gaussian distribution (i.e. the distribution of the solar radiation outside the atmosphere) can be compensated by $W(\lambda)$, but the absorption lines in the far red and near infrared can disturb the color constancy, unless the sensitivity of the eye rapidly decreases there which is definitely not the real situation, (cf. Ref. [10]). Then the only possibility is that there the ratios $P_2(\lambda)/P_0(\lambda)$, $P_1(\lambda)/P_0(\lambda)$ go to constant values. In this case the color coordinates v_i, which are the raw material for the color vision, are insensitive on the part of the distribution which cannot be simulated by protomers. One can, in fact, see this behaviour at the red end of the spectral loop.

At the violet end the color coordinates are less constant. However, the near ultraviolet component of the solar radiation is definitely weaker than the near infrared one. Furthermore, the absorption bands are not so important there, at least above 290 mμ. Anyway, one cannot expect evolutionary gain by producing a proper biochemistry for keeping the proportionality (2.13) with a Gaussian (or any) basis at wavelengths where the sensitivity or the solar radiation is negligible. This may be the reason to see various colors between yellow and red in the near infrared [10].

6. The conformal factor

At the end of Sect. 2 some consequence of the non-uniqueness of the measure of discrepancy was discussed. Ref. [5] shows that this measure Δ is defined only up to a multiplicative constant. But then this constant appears in the distance too:

$$ds^2 = c^2 \frac{1}{H} \frac{\partial^2 H}{\partial p^i \partial p^k} dp^i dp^k. \tag{6.1}$$

Can we fix this factor somehow, or is it a free constant of the theory?

It is easy to see that the value of c does influence the observables. Assume that the true protomeric basis is already known, together with the functions $C(\lambda)$ and $W(\lambda)$. Then, of course, some freedom remains in the description: the functions $P_i(\lambda)$ may be linearly combined with constant coefficients, but, although this does not alter the protomers and the distances, through Eq. (2.6) the color coordinates v_i are generally affected. Thus, if the coordinate system is fixed, the only freedom is a common constant coefficient in all the basic functions, with a proper rescaling in $C(\lambda)$. Now, let us fix this common factor in any convenient way. By preparing the protomers, the distances between these lights can be calculated, and this distance does contain c^2.

This term is a constant conformal factor, and it can be removed by conformal rescaling [13]. However, conformal rescaling is not a coordinate transformation; it connects two different Riemannian spaces between which there is a geometric similarity. Thus the c^2 factor cannot be eliminated without altering the Riemannian space.

In fact, such a factor seems to be fortunate and necessary. Namely, it defines an individual distance scale. Now, by definition, there is a unit distance between just distinguishable colors, and one may expect individual variations in the meaning of the term "just distinguishable", from a house painter to an unskilled person. Such individual differences in normal color vision can be described by the constant conformal factor, and it seems that this is the only possibility to simply describe it. Namely, the metric tensor is homogeneous with zero order in H, so the distance does not change with a constant rescaling in C or W. Similarly, by decreasing the sensitivity factors A_i with a constant, Eq. (2.13) leads to a common constant factor in the P_i's, which is equivalent with a coordinate dilatation in the p^i space, not with a conformal rescaling. Thus c^2 plays a specific role in the formalism, and should not be removed.

While it would be difficult to make a guess for the numerical value of the average of c in the population, something can be told qualitatively. Choose an arbitrary point in either the color or the protomeric space. Then the just distinguishable points lie on a closed curve, whose points are at a unit distance from the central point. However, this distance is measured along geodesic lines, which are generally not straight lines. Thus the closed curves may possess quite complicated shapes. For infinitesimally small distances Eq. (2.14) would give ellipsoids, but small distances could be measured only by means of statistical evaluation of the fluctuations in matching [5], needing long series of matching, and an assumption for the probability distribution.

However, consider Eq. (6.1). If c is sufficiently large, $ds^2 = 1$ for so small dp^i values where the coefficients $H^{-1}(\partial^2 H/\partial p^i \partial p^k)$ are still almost constant, and then the error points lie on an ellipsoid. On the other hand, if c is small, the points of unit distance possess such coordinate values that between them and the central point the metric substantially changes, so then the error curve is not quadratic.

Consequently, if the curve of just distinguishable colors is similar to an ellipse, then c is large. More quantitative statements would need more specification of the particular model.

7. *The minimal analytic model*

We have seen that there are suggestions pointing to a Gaussian basis, at least in the middle of the "visible" range. This is an attractive possibility, presented by the natural selection as a matter of kindness, so let us esteem it, because it enables us to manufacture a simple analytic model. Namely, then

$$P_i(\lambda) = M_{ir}\lambda^r, \tag{7.1}$$

where M_{ik} is a matrix of constant elements, their actual values can be calculated from the shape, orientation and wavelength parametrization of the spectral loop, as it was shown in Sect. 3. Now, $C(\lambda)$ and $W(\lambda)$ are functions of one single peak, so they can be expanded as

$$\begin{pmatrix} C(\lambda) \\ W(\lambda) \end{pmatrix} = \frac{N_K}{\sqrt{2\pi}\,\sigma_K} e^{-\frac{(\lambda-\lambda_K)^2}{2\sigma_K{}^2}} \pi_K(\lambda);\; K = \begin{pmatrix} C \\ W \end{pmatrix}, \tag{7.2}$$

where $\pi_K(\lambda)$ stands for some polynomials. We have seen that C is probably not too far from a Gaussian for wavelengths where Eq. (7.1) can still be used, the approximate parameters are given in Eq. (4.2). The principal illumination is not too Gaussian, however here we are manufacturing the minimal model, so let us substitute both polynomials by constants. Then $W(\lambda) = 1$ can be achieved by means of the gauge transformation (2.16), and the protomers obtain the form

$$V(\lambda) = \frac{N}{\sqrt{2\pi}\,\Sigma} e^{\mp \frac{1}{2\Sigma^2}(\lambda - \Lambda)^2}$$

$$N = (\mp\pi/q_2)^{1/2} e^{q_0 - q_1{}^2/4q_2},$$

$$\Lambda = -\frac{q_1}{2q_2}, \quad \Sigma = \frac{1}{\sqrt{\mp 2q_2}}, \quad q_i = M_{ir}p^r. \tag{7.3}$$

That is, the protomers are Gaussian for the mixtures of spectral colors and white (upper sign), or inverse Gaussians for purples (lower sign). Then the luminance H can be expressed in analytic form as

$$H = \frac{N}{\sqrt{2\pi(\Sigma^2 \pm \sigma^2)}} e^{\mp \frac{1}{2}\frac{(\Lambda - \lambda_0)^2}{\Sigma^2 \pm \sigma^2}}. \tag{7.4}$$

There is a singularity in the extreme purple at $q_2 = 1/(2\sigma^2)$.

Now the metric can be calculated according to Eq. (6.1), using Eqs (7.3–4), but the result is very complicated, so it is more convenient to express the distance in a parametric form. Let us introduce the new variables

$$H' = \ln H(N, \Lambda, \Sigma) = H''(q_0, q_1, q_2),$$

$$X = \frac{\lambda_0\Sigma^2 \pm \Lambda\sigma^2}{\Sigma^2 \pm \sigma^2} = \frac{\lambda_0 + \sigma^2 q_1}{1 - 2\sigma^2 q_2}, \quad Y = \frac{\sigma\Sigma}{\sqrt{2(\Sigma^2 \pm \sigma^2)}} = \frac{\sigma}{\sqrt{2(1 - 2\sigma^2 q_2)}}. \tag{7.5}$$

Then the line element gets the simple form

$$ds^2 = c^2\left[dH'^2 + \frac{1}{2Y^2}(dX^2 + dY^2)\right], \tag{7.6}$$

while the corresponding color coordinates are as follows:

$$v_i = M_{ir}h^r H, \quad h^k = (1, X, X^2 + 2Y^2). \tag{7.7}$$

In the normalized color cordinates x and y the $X =$ const. lines are straight lines converging at the point of the purple divergence, while the $Y =$ const. ones are hyperbolae.

Since the constant matrix M_{ik} is determined by the spectral loop, the metric has three parameters, λ_0, σ and c, of which the first two are parameters of the sensitivity of the eye. The number of the independent Killing vectors [14] is four, namely they are as follows:

$$K_1^i = (1, 0, 0), \quad K_2^i = (0, 1, 0), \quad K_3^i = (0, X, Y), \quad K_4^i = (0, X^2 - Y^2, 2XY). \tag{7.8}$$

The symmetry group is U(1) ⊗ SO(1, 1). The geodesic curves are solutions of the geodesic equation

$$\frac{d^2X^i}{ds^2} + \Gamma_{rs}{}^i \frac{dX^r}{ds}\frac{dX^s}{ds} = 0, \quad X^i = (H', X, Y), \tag{7.9}$$

where s is the path length, and $\Gamma_{ik}{}^{m}$ stands for the Christoffel symbols. As a consequence of the Killing equation [14]

$$K_{i;k}+K_{k;i} = 0 \tag{7.10}$$

(where the semicolon denotes covariant derivative), the quantities of form

$$c_A = K_A^r \dot{X}^s g_{rs}, \qquad \dot{X}^i = \frac{dX^i}{ds} \tag{7.11}$$

are conserved along geodesic curves. Thus it is possible to get first order differential equations for these curves, instead of the second order Eq. (7.9); Eq. (7.11) yields four, while a fifth comes from the definition of s, being $\dot{X}^r\dot{X}_r = 1$, they are as follows:

$$\begin{gathered} \dot{H}' = c_0, \quad \dot{X} = c_1 Y^2, \quad \dot{Y} = (c_2-c_1X)Y, \\ 2(c^{-2}-c_0^2)-c_2^2+c_1c_3 = 0, \\ (X-c_1^{-1}c_2)^2+Y^2 = 2(c^{-2}-c_0^2)c_1^{-2} \quad \text{if} \quad c_1 \neq 0, \\ X = c_3/(2c_2) \quad \text{if} \quad c_1 = 0. \end{gathered} \tag{7.12}$$

The first equation shows that the logarithm of the scalar of luminance changes uniformly on geodesics. Now we can turn to the restricted H = const. problem. The metric on the H = const. surfaces is of constant curvature, and is the same as in Poincaré's half plane model of the hyperbolic (Bolyai) plane. Ref. [15] contains a detailed analysis of this metric.

Eqs (7.12) show that the H = const. geodesics are either semicircles centered on the X axis, or vertical lines. The X axis is the spectral loop (cf. Eq. (7.5)), therefore $Y < 0$ is the region of the imaginary colors which cannot be physically produced as stimuli. At $Y = 0$ the metric is singular, but the curvature is not; starting from any other point, the geodesic distance until $Y = 0$ is infinite.

Here we have not calculated the real C.I.E color coordinates of the points (H', X, Y), because this minimal analytic model is only a rough first approximation, the (7.1) form of the protomeric basis is an approximation, and the error becomes double in the color coordinates containing M_{ik}. Nevertheless, the protomers can be easily identified by their physical parameters N, Λ and Σ. These parameters possess clear meaning, so in this context we will use the expressions hue and saturation as synonims for Λ and Σ, respectively, to avoid lengthy periphrases.

Now, consider a geodesic between two points of different hues. Its length is as follows:

$$\begin{gathered} s_{12}^2 = c^2 \left\{ (H_2'/H_1')^2 + \tfrac{1}{2} \ln^2 \frac{\operatorname{tg} \varphi_2/2}{\operatorname{tg} \varphi_1/2} \right\}, \\ \operatorname{tg} \varphi_{\binom{1}{2}} = - \frac{(X_2-X_1)(Y_2+Y_1 \pm (Y_2-Y_1))}{(Y_2+Y_1)(Y_2-Y_1) \mp (X_2-X_1)^2}, \end{gathered} \tag{7.13}$$

while, for identical hues,

$$s_{12}^2 = c^2\{(H_2'/H_1')^2 + \tfrac{1}{2}\ln^2(Y_2/Y_1)\}. \tag{7.14}$$

If one of Y_1 or Y_2 goes to 0, s_{12} becomes logarithmically infinite. This means that, as we have mentioned, the spectral loop is in the infinity in this model. Such a phenomenon was noted in Ref. [5] based on more general arguments, the fact will be discussed later. Similarly, if the upper point of a vertical geodesic goes to infinity (i.e. to the point of the purple divergence), the distance becomes infinite too.

For the error ellipses, we mentioned in Sect. 6 that they are ellipses only if c is sufficiently large. By assuming this, and remaining on the plane of conscant luminance, Eq. (7.6) leads to

$$\left(\frac{d\Lambda}{\Sigma^2+\sigma^2} - 2\frac{\Lambda-\lambda_0}{(\Sigma^2+\sigma^2)^2}\Sigma d\Sigma\right)^2 + \frac{\sigma^2 d\Sigma^2}{2(\Sigma^2+\sigma^2)^3} = \frac{c^2\Sigma^2}{\sigma^2(\Sigma^2+\sigma^2)}. \tag{7.15}$$

If the widths of the distributions are fixed, one gets

$$d\Lambda = \frac{\Sigma\sqrt{\Sigma^2+\sigma^2}}{2c\sigma} \tag{7.16}$$

for the just distinguishable protomers. Near to the spectral loop it reduces to $d\Lambda = \Sigma/2c$, so the possibility of distinction depends simply on the ratio of the difference of dominant wavelengths and the width, which seems to be realistic. If Λ is fixed, a more complicated formula is obtained, which, for sharp lines, reduces to

$$d\Sigma = \sqrt{2}\frac{\Sigma}{c}. \tag{7.17}$$

Now let us try to explore the manifold. The vertical lines are geodesic curves. The Y axis is the only one along which Λ is constant, $\Lambda = \lambda_0$. Going upward the saturation is decreasing, at $Y = \sigma/\sqrt{2}$ one arrives at the physical white point where $q_i = 0$ (that is, $V(\lambda)$ is constant). Going beyond this point the protomers are inverse Gaussian, so this is the purple region, with increasing saturation. This geodesic terminates at $Y = \infty$, the point of purple divergence. The asymptotic purple line, consisting of mixtures of red and blue spectral lines is not a part of the manifold.

For the other vertical geodesics Λ is moving off λ_0, at $Y = \sigma/\sqrt{2}$ $(\Sigma = \infty)$ $\Lambda = \pm\infty$, above $Y = \sigma/\sqrt{2}$ the light becomes purple, and Λ is returning from the other side of λ_0. The terminal point is again the point of the purple divergence.

Two colors of different hues are connected by a semicircular geodesic curve, centered on the X axis (that is, on the spectral loop). The path length between the two colors depends on the angles of the X axis and the lines pointing to the colors from the center of the semicircle. The curve may or may not protrude into the purple region, depending on the difference of the hues.

The manifold is complete in the sense that its boundaries ($H' = \pm\infty$, $X = \pm\infty$, $Y = 0$ and $Y = +\infty$) are infinitely far from any other point.

Now, we have manufactured an analytic model. It has some distinctive characteristic features, both correct and incorrect ones. The most important such features will be mentioned and briefly discussed here.

The Riemannian space possesses four Killing vectors, it is not of constant curvature, however, it is homogeneous and isotropic. So, in some parameters the ability of color discrimination is uniformly good (or bad) everywhere, but, in the same time, the structure of the space itself offers a natural way to be decomposed into 2+1 dimensions. This natural decomposition seems to be seen, so this model is the most symmetric possible model. The surfaces of constant luminance are of constant curvature, and this seems too handsome. Independently of the facts, it is not obvious, what is the goal of the evolution. On the one hand, it would be profitable to concentrate on certain important parts of the color space. On the other hand, it is quite possible that the important reflectants are more or less uniformly distributed on the color "triangle", at least there does not seem to exist any intimate or necessary correlation between the reflectance function of the surface of an object and its biological benefit. Anyway, note thai if the spectral distribution of the solar electromagnetic radiation were Gaussian, the brain could completely compensate such particularities as the location of its maximum, or its width; in the present minimal analytic model no solar parameter appears. Of course, fundamental laws of physics do not permit a Gaussian solar spectrum, be it ever so good for the color vision. In Sect. 5 we discussed the asymmetry of the Planck spectrum. Let us describe it by a linear function in the principal illumination:

$$W(\lambda) = \frac{N_W}{\sqrt{2\pi}\,\sigma_W}(1+a\lambda)e^{-\frac{(\lambda-\lambda_W)^2}{2\sigma_W^2}}. \tag{7.18}$$

Then the exponential part can again be removed by the gauge transformation (2.16), and then Eq. (2.14) leads to

$$H(N, \Lambda, \Sigma; \lambda_0, \sigma, a) = H(N, \Lambda, \Sigma; \lambda_0, \sigma, 0)\left(1+a\,\frac{\Sigma^2\lambda_0 \pm \sigma^2\Lambda}{\Sigma^2 \pm \sigma^2}\right), \tag{7.19}$$

where $H(N, \Lambda, \Sigma; \lambda_0, \sigma, 0)$ is given by Eq. (7.4). Here one parameter of the Sun explicitly appears, and one cannot expect homogeneous isotropic Riemannian space anymore.

In the present model there is no end of the spectral loop. This phenomenon was discussed in the earlier Sections; there are obviously some cutoffs for the validity of relation (7.1) at two extremal λ values, but such more complicated models will not be discussed here.

It is strange that the spectral loop is infinitely far from the internal points; this seems to be incompatible with the observations. Nevertheless, this prediction is not a consequence of the specific assumptions of the particular model, rather it seems to possess more general reasons. Namely, choose a quite arbitrary but smooth protomeric basis with smooth $C(\lambda)$ and $W(\lambda)$ functions. Take an arbitrary wavelength $\bar{\lambda}$. Then, sufficiently near to $\bar{\lambda}$ the basic functions can be expanded into Taylor series stopping at the quadratic terms, which leads again to Eq. (7.1), only not globally, but locally, and similarly, the constant, linear and quadratic teı ms of the expansions of $C(\lambda)$ and $W(\lambda)$ can be collected into Gaussian functions, with some local values of averages and widths. Then the leading term of H is again of the form (7.4). For very sharp lines at or near to $\bar{\lambda}$ this luminance function is correct, and the metric (7.6) can be obtained. (The calculation is tedious but straightforward.) However, then the spectral loop is again in the infinity. Since here almost nothing has been assumed, it seems that this result is a direct consequence of WT. One may eliminate the problem on a qualitative level by telling that the exact reconstruction of the protomers would require an infinitely large neural network. Since the brain has other things to do, and very sharp protomers are unimportant in the natural environment, being not produced either by illuminations or reflectances, the reconstruction will break down at some surface $\Sigma = \Sigma(N, \Lambda)$.

Finally, the model contains a purple divergence, at an inverse Gaussian protomer with $\Sigma = \sigma$, and, although the manifold is complete as a Riemannian space, the more extreme purples cannot be found anywhere on it. This is a serious defect of the model, but cannot be easily corrected. Extreme purples are composed of extreme blues and reds, as far as possible of sharp lines, but such protomers are improbable. Using extreme inverse Gaussian protomers, they protrude into extremal regions, where the true basis is not Gaussian any more. Thus one cannot expect nice extreme purples in a Gaussian (i.e. analytic) model. Here we wanted to manufacture an analytic model. Probably another analytic model using double Gaussians could be made for purples, but it would not be good for spectral colors, so the advantages would be overcompensated.

8. Anomalies in the color vision

Some 10% of the population possesses some alterations in the color vision, which alterations can be grouped according to various principles. (See e.g. Ref. [7].) The differences, or defects, sometimes are formulated such terms that the person cannot see a color (or rather a pair of colors), or that the color discriminating ability is reduced. Now, WT shows that color vision is a more sophisticated process than a simple limited frequency analysis of the reflected light suggested by Eq. (1.1), either the hardware, or the software can be affected. Here we want to give a brief discussion of the possible defects based on the formalism of WT. This discussion will not be complete, only some characteristic defects will be mentioned. Although until now we have not been using the three receptor model, here we accept that there exist some entities in the eye producing the functions $A_i(\lambda)$.

The usual terminology is not too proper for WT. Namely, e.g. the color blindness indicates that there are some colors which are not seen by the affected observer. Now, consider a protanope. It is told that he cannot see reds and greens, only yellows and blues. Nevertheless, this is a very subjective statement (which is generally not the subjective feeling of the protanope, who learnt that some objects must be red and green, and, of course, not the subjective feeling of the normal trichromat, using the term, but rather the subjective feeling of the normal trichromat about the nature of the subjective feeling of the protanope about colors), so it seems to be better to avoid its use. One may tell that this subjective term is confirmed by unilateral dichromats, but we shall see that the color vision of unilateral dichromats may be a quite complicated process, not characteristic for other dichromats.

WT contains four functions of one variable, $P_i(\lambda)$ and $C(\lambda)$, in addition, our scale--invariant generalization introduces one more, $W(\lambda)$ and a scale constant c. Now, in these models the formation of the color coordinates is according to Eq. (1.1), but the new A_i functions are expressed by P_i and C (cf. Eq. (1.3)). This relation, and some optimal forms of the fundamental functions, are consequences of the natural selection. One may guess that A_i and C belong to the hardware, P_i, W and c to the software. Now, any of the functions can be anomalous. Any specific kind of anomaly leads to a less good color constancy. The main possibilities are as follows:

1. The dimension of the space is reduced to 2 (in classical terms, one kind of receptors is missing). This is color blindness, and this case will be discussed later.

2. The functions A_i are anomalous. Then there are two subcases: either P_i are normal, when relation (1.3) does not hold, WT (or similar theories) cannot be applied, the color discrimination may be good, but the constancy is bad; or P_i are anomalous too, in such a way that the proportionality hold. In this second case the protomeric basis is adapted not to the natural environment, but to the defected hardware. WT can be applied with a proper basis; under special circumstances the color constancy may be (accidentally) good, the error ellipses possess sizes and orientations quite different from usual. Substantial part of anomalous trichromats may belong to Case 2.

3. The function C is anomalous. This is similar to the previous case.

4. The function W is anomalous. Such an anomaly changes the metric and leads to a reduced color constancy, but the negative consequences do not seem too direct, some model calculations would be necessary.

5. The scale constant c possesses an extreme value. If it is extremely large, there is an enormous discrimination ability, which is generally not regarded as an anomaly. If the value of c is very small, the whole color vision is normal, only the ability of color distribution is reduced (e.g. because of lack of experience), and the error "ellipses" have (seemingly) irregular shapes (cf. Sect. 6).

All these particular defects are obvious possibilities, and can easily occur. Ref. [16] gives a detailed comparison of the error ellipses of seven individuals with various types of color vision. One may tentatively set the individuals into the mentioned defect groups as follows:

AB is a normal trichromat. *BC* may be normal too, or may belong to Case 4, because of the slight differences. *CD* seems to belong to Case 5. *DE*, *EF* and *FG* seem to possess the defect 2, with increasing degree, while *GH* is a representative of Case 1. Of course, these comments are illustrations, not the complete analysis of those individual cases.

Since Case 1 is mathematically simplest, we finish the discussion with a model calculation for that case. Consider first a normal set of sensitivity functions A_i compatible with a Gaussian basis

$$A_i(\lambda) = C(\lambda)\,(a_i+b_i\lambda+c_i\lambda^2) \tag{8.1}$$

and remove A_2, with the whole dimension. (If the three kinds of receptors exist indeed, then a definite linear combination of the basic functions is removed.) Now, even then the brain can find such a new luminous efficiency function $\tilde{C}$ that Eq. (1.3) hold and the neutral amplification exist in the protomeric basis

$$Q^R P_R(\lambda) = 1; \quad I = 0, 1 \tag{8.2}$$

and then the second basic function obtains the form

$$P_1 = \frac{a_1 + b_1\varphi + c_1\varphi^2}{Q^R a_R + Q^R b_R \varphi + Q^R c_R \varphi^2}. \tag{8.3}$$

If necessary and profitable, by choosing the ratio Q_1/Q_0 properly, the coefficient of either the linear or the quadratic term in the denominator can be made 0, and this is the maximum which can be done by the brain, since $A_I(\lambda)$ is determined by the hardware. Eliminating the quadratic term, the basis is composed of the functions:

$$P_0 = 1, \quad P_1 = \frac{(\lambda - \lambda_1)(\lambda - \lambda_2)}{\lambda - \lambda_3}. \tag{8.4}$$

Now, the quantities λ_1, λ_2 and λ_3 are determined by the sensitivity functions of the true receptors. Since in WT nothing changes by introducing linear combinations of constant coefficients of the basic functions, the investigation of the metric of the color space of normal trichromats cannot yield any information about the proper linear combinations in the sensitivity functions of the receptors (even the existence of the receptors is not necessary), the key is just the color metrics of dichromats. Therefore again first some analytic model is necessary. Assume then that $\lambda_3 = \lambda_2$; in this case one gets an exponential basis; the protomers have the form

$$V(\lambda) = W(\lambda) e^{q^0 + q^1 \lambda}, \tag{8.5}$$

where again q^I stands for some linear combinations of the protomeric coordinates. In order to manufacture an analytic model, again Gaussian forms will be assumed for $C(\lambda)$ and $W(\lambda)$, by gauge transformations (2.16) $\lambda_C = \lambda_W$ can be achieved, and, mutatis mutandis, the steps of Sect. 7 can be repeated. The protomers are Gaussian functions of fixed width

$$N = \sqrt{2\pi}\, s e^{q^0 + q^1 \lambda_0 + (q^1)^2 s^2},$$

$$\Lambda = q^1 s^2 + \lambda_0, \quad \Sigma = s, \tag{8.6}$$

and the scalar of luminance H gets the form

$$H = \frac{s\sigma}{\sqrt{s^2 + \sigma^2}} \exp\left\{ q^0 + \lambda_0 q^1 + \tfrac{1}{2} \frac{s^2 \sigma^2}{s^2 + \sigma^2} (q^1)^2 \right\}. \tag{8.7}$$

Hence the metric can be calculated, then, introducing $\ln H$ and Λ as new coordinates, the distances are as follows

$$ds^2 = c^2 \left\{ (d \ln H)^2 + \frac{\sigma^2}{s^2} \frac{1}{s^2+\sigma^2} d\Lambda^2 \right\} . \tag{8.8}$$

This is the minimal analytic model for the Riemannian space of a dichromat. Nevertheless, the assumptions made here are more serious than those in Sect. 7; e.g. there is no even suggestion for $\lambda_3 = \lambda_2$. Even in the very fortunate case when such an equality holds when missing one kind of receptors, it will not be true when missing another one. Admitting this, let us investigate the results.

Observe first that the protomers are still Gaussian, only the width is fixed. When $\lambda_2 \neq \lambda_3$, this is not true. Consider then an unilateral dichromat. His protomeric basis (in the brain) will be Gaussian, from experiences with the good eye. Then the data from the defected one will be processed by means of an improper code. Therefore the color naming of unilateral dichromats is not a clear evidence for anything fundamental. Nevertheless, in our special model the dichromat's protomers are among the protomers of the (model) normal population. Using the X, Y coordinates of Sect. 7, $Y = Y_0 = s\sigma/\sqrt{2(s^2+\sigma^2)}$. The $H =$ const. section is a line, on which the spectral loop and the extreme purple line coincide. The physical white point, i.e. $q^1 = 0$ is at λ_0, in the yellow. The color discrimination ability is independent of the wavelength.

This is only a minimal analytic model, with serious neglections and with an ad hoc assumption that two roots coincide. Therefore it is quite possible that this model is irrelevant. Nevertheless, there is a rare color vision deficiency, called tritanopia, which shows some characteristic features similar to the consequences of this analytic model. Tritanopia might be expected as frequent as protanopia and deuteranopia, on the grounds that there are three different possibilities to miss one kind of receptors; in the absence of the "blue" receptors the natural consequence would be a "blue-yellow confusion". Now, first, tritanopia is very rarely observed and, second, it does not seem to be a true color blindness [7]. It is reported that the tritanopic eye of an unilateral tritanope can see approximately normal hues, and these hues are arranged into a normal sequence, nevertheless, with a serious desaturation in the yellow [7].

Now, this is just the effect predicted by the presented analytic approximation. Since the protomers (8.5) are among the protomers (7.3), the optimal protomers for the two eyes coincide. The protomeric form (8.5) enables the brain to evaluate the dominant wavelengths of distributions, and the assumed width is just that of the solar radiation, so there is even a limited color constancy. Nevertheless, the saturation cannot be seen, so some part of the spectral loop must be seen white (or confused with white). Since the maximal sensibility of the eye is in the yellow, a yellow-white confusion is not surprising.

Transforming the metric (8.8) into the cooıdinates of the metric (7.6), it can be seen that protomers of the same H' and X are confused, so on the x, y plane the confused colors lie on straight lines converging at the point of the purple divergence. Nevertheless, this result should not be taken too seriously, being that particular point where the model of Sect. 7 ceases to yield finite results.

Such a reduction of the color space has relatively harmless consequences, so it is expected to be rarely observed, indeed. Of course, the similarity of tritanopia to our model may be quite accidental.

9. *Conclusions*

Weinberg's theory establishes an intimate connection between the metric tensor of the color space and the illuminations and reflectances in the nature, via natural selection. This means an automatic explanation for the fact that (for the overwhelming majority of the population) in this fully subjective space a more or less objective and impersonal metric can exist. In this paper we have shown that WT can be written into a slightly more complicated but scale-invariant form; thus it can be used on arbitrary scale instead of the wavelength. Since the brain probably does not use the notion of physical wavelength for evaluating the data obtained from the eye, the function guaranteeing the scale-invariance (the principal illumination) seems to be necessary indeed, and it is useful for compensating the particularities of the solar spectrum, in order to achieve greater color constancy.

In the generalized formalism a minimal analytic model is given. The metric of this model has an interesting feature: while it is homogeneous, there are four independent Killing vectors with three dimensional transitivity in the three dimensional space, the space is not of constant curvature, and the symmetry structure gives a natural $2+1$ decomposition, according to the experimental facts. Although some other features seem to be unrealistic, this is an interesting explicit example for the specific role of Gaussian bases.

We have also shown that Weinberg's very important results, which originally were formulated for normal and good color vision, remain valid even for some special types of anomalies of the color vision. A theoretical possibility of an interesting type of color blindness (i.e. 2 dimensional color space) is shown, when the dominant hues can be identified, but the saturations cannot. Since the mechanism needs the equality of two (not necessarily related) parameters, such an anomaly is possible but its existence is not necessary.

The author would like to thank Drs I. Borbély and L. Diósi for illuminating discussions.

REFERENCES

[1] H. von Helmholtz, *Sitzber. Preuss. Akad. Wiss.* 1071 (1897).
[2] W. K. Wootters, *Phys. Rev.* **D23**, 351 (1981).
[3] L. Diósi, G. Forgács, B. Lukács, H. L. Frisch, *Phys. Rev.* **A29**, 3343 (1984).
[4] E. Schrödinger, *Ann. Phys.* (Germany) **63**, 397, 481 (1926).
[5] J. W. Weinberg, *Gen. Rel. Grav.* **7**, 135 (1976).
[6] B. Lukács, *Acta Phys. Hung.* **54**, 155 (1983).
[7] G. S. Wasserman, *Color Vision*, J. Wiley Sons, N. Y. 1978.
[8] Landolt-Börnstein, Zahlenwerte und Funktionen, Technik, 3. Teil, Elektrotechnik, Lichttechnik, Röntgentechnik. Springer, Berlin-Göttingen-Heidelberg 1957.
[9] F. Kohlrausch, *Praktische Physik*, 19. Aufl., Teubner, Leipzig 1955.

[10] D. R. Griffin, R. Hubbard, G. Wald, *J. Opt. Soc. Amer.* **37**, 546 (1947).
[11] I. G. H. Ishak, *J. Opt. Soc. Amer.* **42**, 529 (1952).
[12] E. Novotny, *Introduction to Stellar Atmospheres and Interiors*, Oxford Univ. Press, New York-London-Toronto 1973.
[13] G. H. Katzin et al., *J. Math. Phys.* **10**, 617 (1969).
[14] L. P. Eisenhart, *Riemannian Geometry*, Princeton Univ. Press 1950.
[15] H. P. Robertson, T. W. Noonan, *Relativity and Cosmology*, Saunders 1969.
[16] J. A. Van den Akker, J. E. Todd, P. Nolan, W. A. Wink, *J. Opt. Soc. Amer.* **37**, 363 (1947).

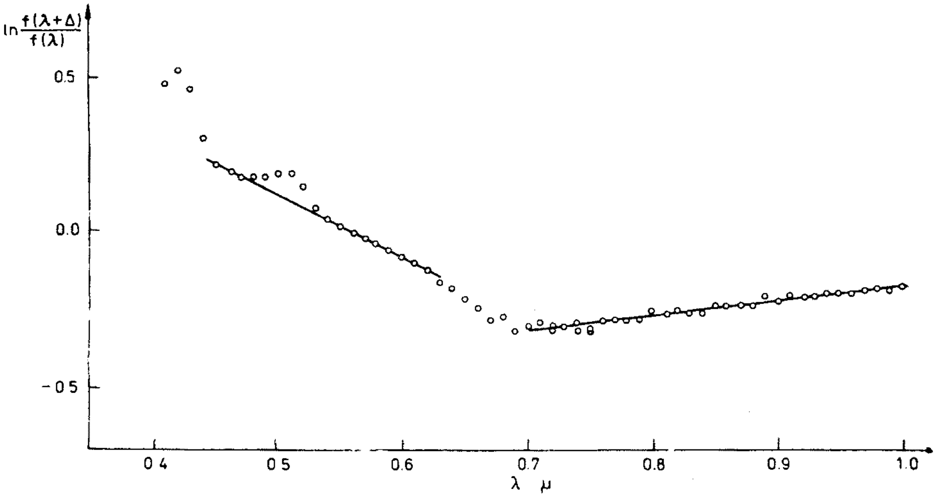

Fig. 2. The relative change of the sensitivity of the eye $f(\lambda)$ [9, 10] in the "visible" and near infrared regions. The displayed quantity is $\ln[f(\lambda+\Delta)/f(\lambda)]$, with $\Delta = 10$ mμ. If f is Gaussian, this curve is linear. More explanations in the text

between extreme blue and extreme red, with

$$\lambda_0 = 555 \text{ m}\mu, \qquad \sigma = 47 \text{ m}\mu \tag{4.2}$$

(although with serious deviation between 475 and 525 mμ). Nevertheless, these parameters should not be taken too seriously; there is some indication that they differ for different populations [11].

One can observe that nothing drastic happens at the red border of the color vision; the decrease of the sensitivity is gradual in the near infrared, and can be even slower than somewhere in the red. On the other hand, Ref. [10] states that in the infrared the color of the sufficiently intensive lights fluctuated between red and yellow. So it seems that there the protomeric basis is already far from Gaussian.

5. *The principal illuminance*

As we have seen, the scale-invariant formalism contains a new function $W(\lambda)$ occurring in all the protomers as a multiplicative factor. In Sect. 2 a suggestion was made that this function represents some preconception about the illumination, when it cannot be observed. Nevertheless, it is necessary to note that we have not proven that W plays this role indeed; in the form of the protomers $W(\lambda)$ appears from merely formal reasons. The formalism itself does not impose any constraint on $W(\lambda)$, e.g. it is even possible that $W(\lambda)$ is constant on the physical wavelength scale, although it would be difficult to find serious arguments for such a hypothesis.

Nevertheless, the formalism definitely needs $W(\lambda)$, and, if a formalism is proper for a problem (which is, of course, an open question for the present case), then its necessary