Universität Bielefeld

Zentrum für interdisziplinäre Forschung

OPTIMAL NONLINEAR CODES

Don McLeod
University of California
San Diego
La Jolla, CA 92093-
0109
USA

Tassilo von der Twer
Bergische Universität
D-42097 Wuppertal
Germany

Perception and the Role of

Evolutionary Internalized Regularities

of the Physical World

1995/96

OPTIMAL NONLINEAR CODES

Don McLeod
University of California
San Diego
La Jolla, CA 92093-
0109
USA

Tassilo von der Twer
Bergische Universität
D-42097 Wuppertal
Germany

Nr. 28/96

Optimal Nonlinear Codes

Don McLeod
University of California
San Diego

Tassilo von der Twer
Bergische Universität
Wuppertal

Introduction

Information theory has for a while been brought to bear on questions about signal processing in the brain (Barlow 61, Atick 92). Economy of coding is thereby linked with ecologically arising probability distributions.

Specifically, the present work addresses the question of gain control: How should an incoming (say, photoreceptor) signal x be recoded for further transport and processing, respecting the needs of a neuron with restricted range and steps of representable values? And how does the induced slicing of the input space ((L,M,S) - cone excitation space or log-cone excitation) look like? A first step in this direction was taken by Laughlin, who realized that in order to have maximal entropy with the distribution of the code values $y(x)$, that is uniform distribution, the function y should run proportionally to an antiderivative of the probability density p of the inputs. But he did not consider noise (Laughlin 81). His result may also be interpreted as yielding the best mutual information between input and output in absence of noise; for the information conveyed from y to x (or vice versa, it is symmetric) is just (H means entropy) $H(x)-H(x|y)=H(y)-H(y|x)$, and y being uniquely determined by x, without noise, we have $H(y|x)=0$, so maximizing $H(y)$ is equivalent to maximizing mutual information.

Atick proposed a two stage model of gain control following some decorrelation procedure and turned successfully to the delicate problem of decorrelation in the presence of noise (Atick 92). He modeled the output as a folding of the input with some propagator (kernel), using a filter for noise suppression. Also, Atick, Li and Redlich put forward a rationale for pruning out directions of decorrelated coordinates, which may be rotated freely with principal component analysis: It is undesirable if one of the channels gets a rather low signal to noise ratio or if the information amount in the channels is too unequal. Choosing an appropriate rotation can remedy this, as was shown for the example of the opponent R−G and the non-opponent R+G channels (Atick et al. 92).

Our focus here is gain control in the presence of noise. Instead of entropy, we consider the criterion of optimal reconstruction of inputs, and we shall obtain a refined

analysis. This will be kept one-dimensional, relying heavily on decorrelation in the sense of Atick being achieved. Decorrelation, and better independence is important for economy of coding as well as for a fine rationale given by Barlow: Associative learning needs knowledge of prior joint distributions, but these joint distributions would carry too heavy a burden if they could not be composed multiplicatively of one-dimensional distributions (Barlow 61). Indeed, some promising work by several people is in progress on generalizing principal component analysis to obtain nonlinear transformations which render the outputs independent in some stronger sense as mere linear independence, while conserving the information of the inputs. Atick (loc. cit.) does so by applying a minimum entropy principle, minimizing more exactly the excess of the sum of output variable entropies over their joint entropy. Moreover, these transformations seem to be even easily performable by networks, as is known for PCA (Parra et al. 95). Assuming this simplification, we can reduce the problems of gain control and of slicing the input space to considering each dimension separately.

We shall derive the nonlinear code functions to have the form

$$y = F^{-1}\left(\alpha\, H + \beta\right), \text{ with } H'\left(x\right) = p^{1/3}\left(x\right), \alpha, \beta \text{ real parameters.}$$

The function F will depend in a simple way on the output noise standard deviation function $\sigma\left(y\right)$. (For the case of the mean squared reconstruction error criterion and without input noise. For the absolue mean error criterion the exponent of p has to be $1/2$, and input noise is no complication in this case.) The slicing of input space goes uniformely proportionally to $p^{-1/3}$ (or $p^{-1/2}$ depending on the criterion). Furthermore, we discuss the task of discriminating inputs instead of identifying them. Finally, we consider and quantify the remarkable advantage of dividing the coding work on several neurons by splitting the input range. This will divide the mean squared errors for n neurons by n^2 instead of dividing only by n by averaging the output values.

1 Mathematical Setting of the Optimal Coding Problems to be Considered

We assume input variables $x_1, \ldots, x_k$ which are already decorrelated. Moreover, we consider them to be even independent. Thus, it will be reasonable to code them by values $y_1\left(x_1\right), \ldots, y_n\left(x_k\right)$, so as to retain independence. This is much easier than the general case $y_i\left(x_1, \ldots, x_k\right), 1 \leq i \leq k$. Not only does it simplify calculations, but also the geometry of slicing the input space. This may now be done in each component separately, thus slicing in a rectangular way by planes orthogonal to the axes of the input variables.

The optimality criterion to be used will be minimization of the squared mean reconstruction error (or absolute mean reconstruction error as well), on some occasions

to be supplemented by consideration of mean firing rates. I.e., given the output values $y_i(x_i)$, $1 \le i \le k$ distorted by independent noises (whose distributions depend, however, on the respective output value), thus given $y_1(x_i) + n_i$, estimate the input values x_i by $\hat{x} = y_i^{-1}(y_i(x_i) + n_i)$ ($y_i^{-1}=$ inverse function of y_i). Now the squared error is $\sum_i (x_i - \hat{x}_i)^2$, and the mean squared error is obtained by taking the expected value of this variable under the joint distribution of the input variables x_i and the error variables n_i. Choose the coding functions $y_i(x_i)$ so as to minimize this.

Our problem may be greatly simplified by using that the noises are independent of one another and that y_i only depends on x_i. Namely, the $2k$-dimensional integral to be minimized can be split in a sum of 2-dimensional integrals (with variables x_i, n_i, respectively), see Appendix C for the argument.

So we are left with the following problem:

Given a single input variable x (taking, e.g., values on some finite interval) and its probability density $p(x)$, construct a coding function $y(x)$ taking for simple idealization any real value in the appropriate range of firing rates for neurons, so that the mean squared reconstruction error for input x, given output plus noise, will be . minimal.

Now, the mean has to be taken with respect to the joint distribution of input and error, which may be written $p(x) \cdot q(n \mid y(x))$, with the conditional (on $y(x)$) density q for the noise variable. So we have the variational problem:

$\int dx\, p(x) \int dn\, (x - y^{-1}(y(x) + n))^2\, q(n \mid y(x)) \to$ minimum, constrained by $y(a) = 0$, $y(b) = \max(> 0)$, x ranging from a to b.

To achieve a more amenable ordinary variational problem, we adopt the following two simplifications:

- Apply linear approximation to the value $y^{-1}(y(x) + n)$, taking y^{-1} to be differentiable everywhere, thus replace $y^{-1}(y(x) + n)$ by $x + \frac{n}{y'(x)}$. (See the remark at the end of this section to justify this.)
  So the squared reconstruction error at x becomes $\frac{n^2}{y'^2(x)}$.

- The next simplification is now exact again:
  Replace $n^2 / y'^2(x)$ by $\sigma^2(y(x)) / y'^2(x)$ and take the value n^2 to be the constant variance $\sigma^2(y(x))$ of the noise distribution conditional to $y(x)$, to be occurring with probability 1, thus getting rid of integrating over the distribution $q(n \mid y(x))$. Here is the reason:

$$\int \frac{n^2}{y'^2(x)} q(n \mid y(x))\, dn = \frac{1}{y'^2(x)} \cdot \int n^2 q(n \mid y(x))\, dn = \frac{1}{y'^2(x)} \cdot \sigma^2(y(x)),$$

since the noise conditional to $y(x)$ has mean 0, and so the expected value of n^2 is just the variance. (For the absolute mean error criterion this does not work so nicely.)

3

Our variational problem has become simply:

Given a distribution $p(x)$, find a strictly monotonic differentiable code $y(x)$ satisfying:

(i) $\text{MSE}(y) = \int_a^b dx\, p(x) \frac{\sigma^2(y(x))}{y'^2(x)} \to minimum.$
    (MSE in short for 'mean squared error'.)

(ii) $y(a) = 0$, $y(b) = \max$ (boundary conditions).

This problem will completely be solved, see section 2.4 for the general solution, and Appendix A for the completion of the reasoning.

We shall provide the optimal codes y ('pleistochromes'), the optimal attainable MSE as well as the completely general formula for the thickness of slices in input space:

$$\frac{\sigma(y(x))}{|y'(x)|} \text{ is proportional to } p^{-1/3}(x).$$

Sections 2.1–2.3 will give the specialization of the general results to three forms of $\sigma(y(x))$: $\sigma(y(x)) = \text{constant}$, $\sigma(y(x)) = \sqrt{y(x)}$ $(y(x) \geq 0)$, and $\sigma(y(x)) = \sqrt{y(x) + \sigma^2}$, σ^2 a constant. The second is motivated by considering $y(x)$ as the expected value of a Poisson variable, the third respects the fact that there is some residual activity in neuron cells.

Section 3. exploits the results of section 2. for a treatment of split codes, and section 4. will reduce the problem of optimal codes for discerning input-pairs (e.g., colors at edges) to that of optimal codes for reconstructing a single input value. Input noise will be considered in section 5.

We close this introduction by the announced remark on the adopted linear approximation. This might seem somewhat reckless, $\sigma(y(x))$ being not so tiny as an usual Δy should be for linear approximation to work. But we may give two reasons to justify linearization (other than simplification, of course), one quantitative and one with regard to the intended application.

Quantitative: Linearization, especially bad one, tends to exaggerate the errors. So, a good code with small linearized mean error will turn out even better. On the other hand, examples show that nonlinearized errors usually should be of the same order of magnitude as the corresponding linearized ones. So corrections in the pleistochromes for true errors will presumably be of little importance while providing considerable difficulties.

Application-directed: Consider a number of neurons, each firing with mean value $y(x)$ and (output) error variance $\sigma^2(y(x))$, working together to give an average value. Then the expected value of this average is again $y(x)$, but the variance of the average is only $\sigma^2(y(x))$ divided by the number n of neurons. This may now well be tiny enough to allow for linearization. But the MSE-integral is then only altered by the factor $1/n$, thus leaving the variational problem unchanged. And a single of the neurons should have the same expected value $y(x)$ as the ensemble by its average.

4

2 Optimal Nonlinear Coding by Firing Rates of a Single Neuron, Using Strictly Monotonic Functions

Throughout this section, we require the coding functions $y(x)$ to be strictly monotonic (in order to be able to invert them) and, moreover, differentiable (for linearization of the inverse). With respect to the foremost intended application to color coding the optimal codes $y(x)$ will be called 'pleistochromes' – they allow for distinguishing many colors, remembering Euler's 'brachystochrones', the first solution of a significant variational problem.

We shall use the criterion of minimal mean squared error, but indicate the changes necessary for the absolute mean error criterion.

Some technical prerequisites are assumed, including some notation and terminology:

x: Input variable; integrals are always over the range of x, unless indicated otherwise.

p(x): Probability density of x; p is assumed to have but finitely many zeroes in the range of x. Furthermore, $\int_{\text{range of } x} p^{1/3}(x)\, dx$ shall have a finite value. (This will be satisfied if p is any interesting distribution. Problems will only occur if the x-range is infinite and p not approaching zero fast enough, e.g., with $p(x) \sim 1/x^3$.)

H(x): The antiderivative of $p^{1/3}$ ranging from 0 to some maximal value. 'Antiderivative' of $p^{1/3}$ means: $H'(x) = p^{1/3}(x)$ for all x in the x-domain. ('Primitive' is also used for this, but with entirely different meaning, too. 'Integral' is equivocal and thus avoided.)

y(x): (Sometimes with parameter-indexes.) Some optimal code, 'pleistochrome'.
We shall indicate how to adjust parameters for some suitable behavior, e.g. range. In principle, they could take only positive values, only negative, or both: But sometimes we shall for simplicity only consider nonnegative ones. There will be rising as well as falling examples, and a rationale for choice will be given.

2.1 The case of constant noise variance

Let the noise variance be the constant $\sigma^2 > 0$.

a) The pleistochromes are

$$y_{\alpha,\beta}(x) = \alpha H(x) + \beta,\ \alpha \neq 0,\ \beta \text{ arbitrary.}$$

If $y_{\alpha,\beta}$ is required to be rising from 0 to some maximal firing rate max, then α and β become fixed. (Remember that H is a specific antiderivative of $p^{1/3}$.) Letting x range from a to b, the equations to be used are ($H(a) = 0, H(b) > 0$):

$$0 = \alpha \cdot 0 + \beta, \text{ so } \beta = 0.$$
$$\text{max} = \alpha \cdot H(b), \text{ so } \alpha = \text{max}/H(b).$$

Falling y requires choosing $\alpha < 0$. $\alpha = -\text{max}/H(b)$, $\beta = \text{max}$ gives the same range as above and thus (see b)) the same MSE. So it turns out that the reconstruction error criterion alone cannot decide which direction to take. But for skewed p the further criterion of minimizing the mean firing rate, i.e., $\int y(x) p(x) dx$ ($y(x) \geq 0$), will provide a decision: Use small firing rates where p is large. The same applies for the other cases of noise.

b) The mean squared error is

$$\text{MSE } (y_{\alpha,\beta}) = \int \frac{\sigma^2}{y'^2(x)} p(x) dx = \frac{\sigma^2}{\alpha^2} \int p^{1/3}(x) dx.$$

c) The corresponding thickness of slices in input space is

$$\frac{\sigma}{|\alpha|} p^{-1/3}(x) \text{ at point } x.$$

Using the absolute error criterion $ME(y) = \int \frac{\sigma}{|y'(x)|} p(x) dx \to$ min. gives completely analogous results, only $p^{1/3}$ has everywhere to be replaced by $p^{1/2}$.

2.2 The case of Poisson noise

In order to avoid lousy distinctions of completely analogous cases, we only consider nonnegative codes $y(x)$. The Poisson noise output variance is then simply $y(x)$, the same as the expected output neuron firing rate.

a) The pleistochromes are

$$y_{\alpha,\beta}(x) = (\alpha H(x) + \beta)^2, \ \alpha \neq 0, \ \beta \text{ arbitrary}.$$

The parameters α and β are to be restricted so that $\alpha H(x) + \beta \geq 0$ for all x or $\alpha H(x) + \beta \leq 0$ for all x. Otherwise, $y_{\alpha,\beta}$ would not be monotonic. Qualitatively, we have ascending or descending (for Gaussian-shaped p: sigmoid) curves as in the constant variance case. The occurring U-shaped curves just excluded for the purposes pursued here are in fact useful, cf. section 3.

b) The mean squared error is

$$\text{MSE } (y_{\alpha,\beta}) = \int \frac{y_{\alpha,\beta}(x)}{y'^2(x)} p(x) dx = \frac{1}{4\alpha^2} \int p^{1/3}(x) dx.$$

6

c) The corresponding thickness of slices in input space is

$$\frac{1}{4\,|\,\alpha\,|}\,p^{-1/3}\,(x) \quad \text{at point } x.$$

The striking similarities to 2.1 will be explained in 2.4. Again, replace 1/3 by 1/2 for the absolute error condition.

Here we present for the case of (one and the same) normal distribution a graphical display of Laughlin's curve (the steeper of the crossing curves) compared to our pleistochromes for constant noise variance (the other of these) and for the Poisson noise (on the right); each curve uses the same output range:
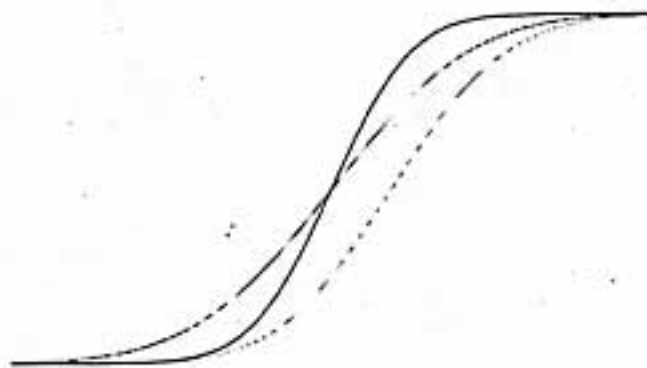

Figure 1: Laughlin´s curve (steepest) compared to pleistochromes for two noise cases, all for the same Gaussian input distribution

Thus for constant noise variance we have a similar curve to Laughlin's, only less steep. (It is the same as Laughlin's would give for three times the standard deviation.) But for Poisson noise, the point of inflection moves to the right.

2.3   The case of hybrid noise

Again, we only consider nonnegative codes $y\,(x)$. The hybrid (modified Poisson) output noise variance is $y\,(x) + \sigma^2$, if $y\,(x)$ is the expected output value, $\sigma^2 > 0$.

a) The pleistochromes are

$$y_{\alpha,\beta}\,(x) = (\alpha\,H\,(x) + \beta)^2 - \sigma^2, \; \alpha \neq 0, \; \beta \text{ arbitrary.}$$

But the parameters α, β have to be confined so that $\alpha\,H\,(x) + \beta \geq \sigma$ for all x or $\alpha\,H\,(x) + \beta \leq \sigma$ for all x, so that $y_{\alpha,\beta}$ be monotonic and positive.
The resulting curves are very similar to those of 2.2, but we have no U-shaped curves taking only nonnegative values. For the U-shaped $(\alpha\,H\,(x) + \beta)^2$ have a zero, so subtracting σ^2 gives negative values. This is also meaningful, cf. section 3.

7

b) The mean squared error is again $\frac{1}{4\alpha^2}\int p^{1/3}(x)\,dx$.

This is exactly the same as in 2.2. But note that for y in the same range we need a smaller α here than in the Poisson case, hence have a greater reconstruction error as obviously to be expected by the raised output variance.

c) The thickness of slices remains the same as in 2.2.

2.4   The case of general $\sigma\left(y\left(x\right)\right)$

Let $\sigma\left(y\right)$ be any function taking only nonnegative values, associating a standard deviation to the output value y. Furthermore, we require the function $\frac{1}{\sigma(y)}$ to have a finite integral over the range of output values y. In Appendix A, 2., we shall establish the following differential equation for the pleistochromes y, given output noise variance $\sigma^2\left(y\left(x\right)\right)$ for expected output value $y\left(x\right)$ with $\alpha \neq 0$ so as to render y constant:

$$\frac{y'\left(x\right)}{\sigma\left(y\left(x\right)\right)} = \alpha\, p^{1/3}. \tag{1}$$

(Thus always thickness $= \dfrac{\sigma\left(y\left(x\right)\right)}{\left|y'\left(x\right)\right|} \sim p^{-1/3}\left(x\right)$.)

This equation is easily solved for y by separation of variables: We have

$$\int_a^x \frac{y'\left(x\right)}{\sigma\left(y\left(x\right)\right)}\,dx = \int_a^x \alpha\, p^{1/3}\left(x\right)\,dx = \alpha\,H\left(x\right).$$

Moreover,

$$\int_a^x \frac{y'\left(x\right)}{\sigma\left(y\left(x\right)\right)}\,dx = \int_{y(a)}^{y(x)} \frac{1}{\sigma\left(u\right)}\,du = F\left(y\left(x\right)\right) - F\left(y\left(a\right)\right),$$

with an antiderivative F of the function $\frac{1}{\sigma}$, so $F'\left(u\right) = \frac{1}{\sigma(u)}$. Thus, $F\left(y\left(x\right)\right) = \alpha H\left(x\right) - F\left(y\left(a\right)\right)$. So, since F has an inverse F^{-1}, and calling $\beta = -F\left(y\left(a\right)\right)$:

$$y_{\alpha,\beta}\left(x\right) = F^{-1}\left(\alpha\,H\left(x\right) + \beta\right). \tag{2}$$

(Thus, all pleistochromes are functions of $\alpha\,H\left(x\right) + \beta$.)

This is the general formula from which all pleistochromes may be derived, given a formula for $\sigma\left(y\right)$. For instance, with Poisson noise:

$$\sigma\left(y\right) = \sqrt{y}.$$

8

We have $\int \frac{1}{\sqrt{y}} dy = 2\sqrt{y} + c$ (indefinite integral), so $F(y) = 2\sqrt{y}$ will do. This gives the pleistochromes $y_{\alpha,\beta}(x) = F^{-1}(\alpha H(x) + \beta) = \frac{1}{4}(\alpha H(x) + \beta)^2$ (parametrization slightly different from 2.2). Furthermore, we can derive from (1) the general formula for the MSE attained by the pleistochrome $y_{\alpha,\beta}$ of formula (2):

$$MSE(y_{\alpha,\beta}) = \frac{1}{\alpha^2} \int_a^b p^{1/3}(x) \, dx. \qquad (3)$$

Remember that $\text{MSE}(y_{\alpha,\beta}) = \int_a^b \frac{\sigma^2(y(x))}{y'^2(x)} p(x) \, dx$ and that by (1) $\frac{\sigma^2(y(x))}{y'^2(x)} = \frac{1}{\alpha^2} \cdot p^{-2/3}(x)$.

2.5 Remark on transformations caused by transforming the input variable

Introduce the new input variable $\tilde{x}$ by $\tilde{x} = f(x)$, f differentiable and everywhere $f' > 0$ (or, alternatively, $f' < 0$). For instance, one might wish to consider log cone excitation instead of cone excitation. We want to relate the according pleistochromes $\tilde{y}(\tilde{x})$ to those for $x, y(x)$. We do not get $\tilde{y}$ from a pleistochrome y for input x by simply setting $\tilde{y}(\tilde{x}) = y(x)$. Instead, we have first to transform the probability distribution $p(x)$ to that of $\tilde{x}$:

$$\tilde{p}(\tilde{x}) = (f^{-1})'(\tilde{x}) \cdot p(f^{-1}(\tilde{x})).$$

Now, we get the analogous results, only for $\tilde{p}$ instead of p. Expressed in x and p, the result becomes

$$\tilde{y}(\tilde{x}) = \tilde{y}(f(x)) = F^{-1}(\alpha K(x) + \beta), \text{ with } K'(x) = p^{1/3}(x) \cdot f'^{(2/3)}(x),$$

where F is an antiderivative of the function $1/\sigma(\tilde{y})$.
But the expression in $\tilde{x}, \tilde{p}$ is again simply

$$\tilde{y}(\tilde{x}) = F^{-1}\left(\alpha \tilde{H}(\tilde{x}) + \beta\right), \quad \tilde{H}'(\tilde{x}) = \tilde{p}^{1/3}(\tilde{x}),$$

which will be more convenient especially when considering input noise in addition, cf. section 5.

3 Splitting the Range of Input Values and Coding the Parts by the Full Firing Range of Different Neurons

So far, we used only strictly monotonic output functions for coding the full input range by a single neuron. Distribution of the coding word on more neurons, each

responsible for a part of the input range, has advantages over averaging by several neurons to be pointed out.

First, we describe the procedure of splitting in two parts, using just two neurons. Split the range of x in two parts, one from a to x_0, one from x_0 to b. (The choice of x_0 will be optimized later on.) Associate coding functions $y_1(x)$, $y_2(x)$ to each neuron with the properties: $y_1(x) = 0$ for $x \geq x_0$, $y_2(x) = 0$ for $x \leq x_0$, whereas y_1 shows on $[a, x_0[$ the full range of firing rates, y_2 does the same on $]x_0, b]$, and these functions shall be strictly monotonic on their interesting intervals, as before. Now, reconstruction of x-value is done by first deciding which of the two neurons is firing and then use its monotonicity. So we shift from, e.g.,
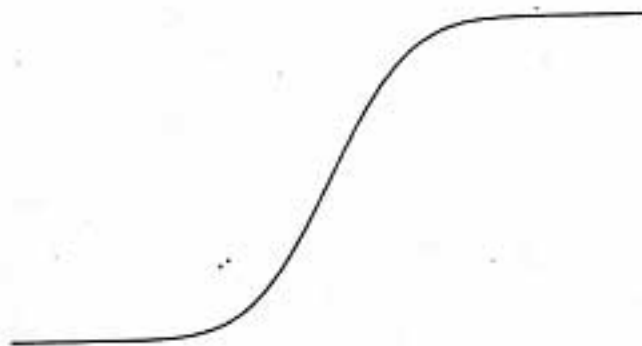

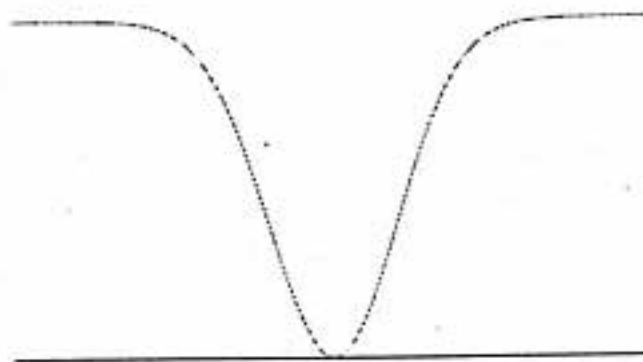Figure 2: Typical sigmoidal nonsplit code

to

Figure 3: Corresponding split code for Poisson noise, using the same range

We want to know the optimal shape of both functions y_1, y_2. It will be convenient to combine both arcs by $y = y_1 + y_2$. Then we have

$$\text{MSE (split code)} = \int_a^{x_0} \frac{y_1(x)}{y_1'^2(x)} p(x)\, dx + \int_{x_0}^b \frac{y_2(x)}{y_2'^2(x)} p(x)\, dx = \int_a^b \frac{y(x)}{y'^2(x)} p(x)\, dx,$$

10

for Poisson noise. (Note that y is no longer monotonic.) Minimizing both sum-mands amounts to the same as minimizing the right-hand integral which we shall call $MSE(y)$.

To do this, we may go either way, the answer will turn out the same: Use the solutions of 2.2 for y_1 on $[a, x_0]$ and for y_2 on $[x_0, b]$ separately. They remain the same: Taking p on a restricted x-interval does not change anything, for the calculations only use the fact that p be a positive function. Or take the (possibly skewed) U-shaped extremal curves for the whole y, which were dismissed in 2.2.

Of course, there remain choices concerning the rising or falling of each arc. But for usual one-peaked probability distributions p the U-shaped combination will have to be performed with respect to the second criterion of minimal mean firing rates. In the other two noise cases, we have, assuming $p(x_0) \neq 0$, the bottom:
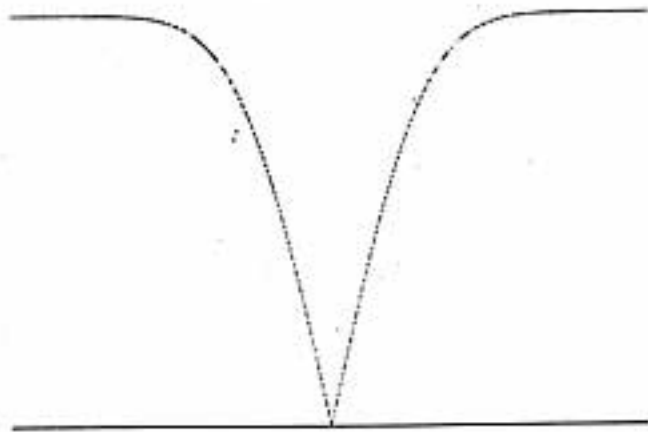

Figure 4: Split code with hybrid noise

Thus, these noise forms require a one-sided jumping derivative at x_0 where $y(x_0) = 0$, which is reasonable.

If $p(x_0)$ happens to be zero, then we get U-shape also in these cases which, however, do not stem from global solutions. But for U-shaped p, for instance, the firing rate criterion would favor a shape like λ for the split code.

We summarize the cost and gain by the splitting procedure:

- Doubling the number of neurons, but without raising the mean firing rate. On the contrary, this will be considerably diminished for usual one-peaked distrib-utions.

- Pushing down the MSE by a factor 4, in all cases of output noise considered here. This is achieved by splitting at the point x_0 where the value of the involved antiderivative of $\alpha p^{1/2}$ (this ranging from 0 to max) is max/2. This choice of the splitting point is optimal, irrespective of whether p is symmetric or not.

11

This is a good result to be compared to what is achieved and lost by taking two neurons both giving a noisy unitary code and reducing MSE by averaging:

- Reduction of MSE by by only a factor 2.

- Doubling the number of neurons.

So the semi-digitalizing by splitting up has advantages. But the procedure cannot go on very far due to aliasing problems especially caused by input noise. Thus, averaging will still be useful.

For the simple case of uniform distribution p, the reduction of MSE by a factor 4 can easily be explained (for the simple general argument, see Appendix B). The x-interval is halved by x_0. The ascending arc is just given by quenching the standard unitary pleistochrome with a factor 2. So the slope is doubled, the values remain the same. According to the formulas for the pleistochromes, this gives a factor $\frac{1}{4}$ for MSE. In addition, the interval is halved, another factor $\frac{1}{2}$. For the two arcs together we get thus a factor $\frac{1}{4}$ for MSE, since their parts in the total MSE are equal.

4 Considering the Criterion of Optimal Discrimination of Pairs of Input Values

We shall argue that this does not change the problem at all. Suppose a joint distribution $p(x_1, x_2)$ of pairs of input values to be given. (The intended application is to color edges.) The problem of pair discrimination may now be posed as follows:
Do the coding $y(x)$ so that the reconstruction error of the differences $x_1 - x_2$ becomes minimal. (We shall use the absolute error criterion here in order to simplify the problem.) This means for Poisson noise and $y' > 0$:

$$\int \int dx_1 \, dx_2 \left(\frac{\sqrt{y(x_1)}}{y'(x_1)} + \frac{\sqrt{y(x_2)}}{y'(x_2)} \right) p(x_1, x_2) \rightarrow \min.$$

(Side condition to be added as above.) Abbreviate $z(x) = \frac{\sqrt{y(x)}}{y'(x)}$. Now, the double integral can be disentangled to give

$$\int \int dx_1 \, dx_2 \, z(x_1) \, p(x_1, x_2) + \int \int dx_1 \, dx_2 \, z(x_2) \, p(x_1 x_2).$$

But the first summand is

$$\int dx_1 \, z(x_1) \int dx_2 \, p(x_1, x_2) \;=\; \int dx_1 \, z(x_1) \, p_1(x_1),$$

$$\text{with} \quad p_1(x_1) \;=\; \int dx_2 \, p(x_1, x_2) = \text{marginal distribution of } x_1.$$

Doing the same for the second summand, we end up with the problem solved above for a one-dimensional distribution. Notice that we do not use any assumption on the independence of the pairs (x_1, x_2), which would of course not be realistic.

For the mean squared error criterion there is no similar reduction, due to the mixed term then appearing.

5 Optimal Nonlinear Codes for Additional Input Noise

We apply an input noise to input x with variance $\sigma_1^2(x)$, thus possibly dependent on x, leaving the output noise with variance $\sigma^2(y(x))$ as before.

After linearizing and taking the mean with respect to the distributions of both noise variables, the squared error becomes $(\sigma_1(x) + \sigma(y(x))/y'(x))^2$.

So we have the mean squared reconstruction error

$$MSE(y) = \int d\dot{x}\, p(x) \left(\sigma_1(x) + \frac{\sigma(y(x))}{y'(x)}\right)^2.$$

The variational problem for y to minimize this amounts to the differential equation (cf. Appendix A, 2.)

$$\frac{y'}{\sigma(y)} = \frac{\alpha p^{1/3}(x)}{1 - \sigma_1(x)\,\alpha p^{1/3}(x)}.$$

The solutions are:

$$y_{\alpha,\beta}(x) = F^{-1}(K_\alpha)(x) + \beta, \quad \alpha \neq 0, \quad \text{and for } \alpha > 0: 1 - \sigma_1 \alpha p^{1/3} \geq \varepsilon > 0, \ \beta \text{ arbitrary,}$$

with K_α (quite different from the former αH) an antiderivative of $\alpha p^{1/3}/(1 - \sigma_1 \alpha p^{1/3})$, F an antiderivative of $1/\sigma$. Note the serious restriction for ascending pleistodromes $(\alpha > 0)$ to occur; it is due to the presence of input noise. Constant input noise gives no essential simplification. But the formulae for MSE and thickness of slices in input space remain the same – thickness at x is $\sigma_1(x) + \sigma(y(x))/y'(x) \sim p^{-1/3}(x)$.

Curiously, using the absolute error criterion for the added input noise case does not provide any change if we put each term of the linearized error sum in absolute bars as usual. But this is less exact, also due to the replacing of the noise variables by their standard deviations not being not quite exact (cf. section 1.)

13

A Appendix A: Calculations of the Pleistochromes

A.1 The case of constant output noise variance by using ordinary differential calculus, and a concrete fashion of generalizing to any variance function $\sigma^2(y)$

Our variational problem reads

$$MSE(y) = \int_a^b \frac{\sigma^2}{y'^2(x)} p(x)\, dx \rightarrow \min., \text{ given } y(a) < y(b).$$

This is simple, since the integrand does not contain y. So we may easily first find y', then y by the boundary values. We assume $p > 0$ throughout the x-range, up to finitely many points which may easily be avoided in the $p(x_i)$ to be mentioned below. Suppose we want to know the number $z_i = y'(x_i)$ for an optimal y just for a finite number of values x_i, say, evenly distributed in $[a, b]$. For a monotonically strictly increasing y we seek only $z_i > 0$. Taking many values, we have with $p_i = p(x_i)$:

$$\int_a^b \frac{\sigma^2}{y'^2(x)} p(x)\, dx \approx \sum \frac{\sigma^2}{z_i^2} p_i \Delta x_i, \text{ and } \int_a^b y'(x)\, dx \approx \sum z_i \Delta x_i,$$

with arbitrary precision. With $\Delta x_i = const.$ value we have the problem:

$$\text{Minimize } \sum \frac{\sigma^2}{z_i^2} p_i \quad \text{subject to the conditions } \sum z_i = c > 0,\; z_i > 0.$$

This is a standard extremum problem solved by writing the partial derivatives and introducing a Lagrange multiplier:

$$\frac{\partial}{\partial z_i}\left(\sum \frac{\sigma^2}{z_i^2} p_i - \lambda\left(\sum z_i - c\right)\right) = 0, \text{ as well as } \frac{\partial}{\partial \lambda}() = 0.$$

So we have $\frac{-2\sigma^2 p_i}{z_i^3} - \lambda = 0$ for all i, and end up with

$$z_i = \left(-\frac{\lambda}{2\sigma^2}\right)^{1/3} \cdot p_i^{1/3} = \alpha \cdot p_i^{1/3},\; \sigma > 0,$$

α being fixed by the side condition.

This gives indeed not only local, but also global minimum for our standard extremum problem; the 'global' part is to be seen as follows: There are fixed numbers ε, δ with

14

$0 < \varepsilon < z_i < \delta$, for any solution (z_i) of the problem. (If any z_i is too near to zero, then $\Sigma \frac{\sigma^2}{z_i^2} p_i$ becomes too large, and for any z_i too great Σz_i becomes too great.)

So our function $f(z_1, \ldots, z_n) = \Sigma \frac{\sigma^2}{z_i^2} p_i$ is restricted to the compact set of vectors (z_i), $\varepsilon < z_i < \delta$ for all i and $\Sigma z_i = c$. Thus f has a global minimum on this set, and this must be the local minimum just found, since it cannot be on the border of the set.

From the solution $z_i = \alpha p^{1/3}(x)$ it is easily read off that one should have $y'(x) = \alpha p^{1/3}(x)$ for all x in order to minimize the MSE for a fixed value of $\int y'(x)\,dx$, requiring y' to be strictly positive up to finitely many points. So we get the pleistochromes $y_{\alpha,\beta}(x) = \alpha H(x) + \beta$, with $H'(x) = p^{1/3}(x)$. Moreover, we can show that these give, for $\alpha > 0$, absolute minima for our problem, in the class of function y with $y' > 0$ up to finitely may points:

Take a pleistochrome $\bar{y}(x)$, so $\bar{y}'(x) = \alpha p^{1/3}(x)$, $\alpha > 0$, and $\int \bar{y}'(x)\,dx = c > 0$. Assume one had a function $y' > 0$ up to finitely many points with $\int \frac{\sigma^2}{y'^2(x)} p(x)\,dx < \int \frac{\sigma^2}{\bar{y}'^2(x)} p(x)\,dx$, $\int y'(x)\,dx = c$. For fine enough partitions of the integration interval by intermediate values x_i (with $y'(x_i), \bar{y}'(x_i) > 0$ always) we have

$$* \quad \sum \frac{\sigma^2}{y'^2(x_i)} p(x_i)\,\Delta x_i + \varepsilon < \sum \frac{\sigma^2}{\bar{y}'^2(x_i)} p(x_i)\,\Delta x_i \text{ , with a fixed } \varepsilon > 0,$$

due to the inequality assumed. Also $\sum \bar{y}'(x_i)\,\Delta x_i$ and $\sum y'(x_i)\,\Delta x_i$ are both arbitrarily close to c. Since $\bar{y}'(x_i) = \alpha \cdot p^{1/3}(x_i)$, $z_i = \bar{y}'(x_i)$ is arbitrarily close to the solution of the minimum problem

$$\sum \frac{\sigma^2}{z_i^2} \Delta x_i \to \min, \text{ with the condition } \sum z_i\,\Delta x_i = \sum y'(x_i)\,\Delta x_i.$$

Thus * is impossible.

Having done the simplest case of constant noise variance, we can easily generalize to arbitrary variance functions whose reciprocal be integrable in the domain considered.

The problem is to minimize $\int_a^b \frac{\sigma^2(z(x))}{z'^2(x)} p(x)\,dx$ for strictly monotonic differentiable z, given $z(a)$ and $z(b)$.

We look for a transformation $y = f(z)$ so that the integral to be minimized carries over to $\int_a^b \frac{1}{y'^2(x)} p(x)$ which we already know to minimize.

For this purpose, we should have

$$\frac{\sigma(z(x))}{z'(x)} = \frac{1}{f'(z(x)) \cdot z'(x)} \text{ , thus } f'(z(x)) = \frac{1}{\sigma(z(x))} .$$

So we choose f to be any antiderivative of $\frac{1}{\sigma(z)}$. Indeed, f is one-to-one, and so starting with any strictly monotonic code z and values $z(a), z(b)$, define $y = f(z)$.

y is again strictly monotonic and differentiable, its values $y(a)$ and $y(b)$ being fixed by $z(a), z(b)$, and vice versa. Moreover,

$$\int_a^b \frac{\sigma^2(z(x))}{z'^2(x)} p(x)\, dx = \int_a^b \frac{1}{y'^2(x)}\, dx.$$

If the former is minimal for given $z(a)$ and $z(b)$, then the latter is for the corresponding given $y(a)$ and $y(b)$, and vice versa.

So the pleistochromes for the variance function $\sigma^2(z)$ are exactly the pleistochromes for constant noise variance transformed by f^{-1}, i.e., $z_{\alpha,\beta}(x) = f^{-1}(\alpha H(x) + \beta)$.

A.2 Treating the general case by variational calculus

We have to minimize $\int_a^b \frac{\sigma^2(y(x))}{y'^2(x)} p(x)\, dx$ for differentiable y with, say, $y' > 0$ but for finitely many values, given $y(a)$ and $y(b)$, $y(a) < y(b)$. Formally, this is more serious, since y and y' occur both. The immediate Euler–Lagrange equation looks fierce. But the substitution $z(x) := \frac{\sigma(z(x))}{y'(x)}$ will help. Now we have to minimize $\int_a^b z^2(x) p(x)\, dx$ with the appropriate constraint $\int_a^b \frac{1}{z(x)} dx = c > 0$. (Notice that the constraint amounts to the same as prescribing boundary values for y, by a change of variable.) This is simple. Introducing a Lagrange multiplier λ, we have

$$F(x,z) = z^2 p(x) - \frac{\lambda}{z}.$$

The corresponding Euler–Lagrange equation is

$$\frac{\partial F}{\partial z}(x,z) = 2zp(x) + \frac{\lambda}{z^2},$$

so $z = z(x) = \beta \cdot p^{-1/3}$, with a $\beta > 0$, the formula for thickness of slices. (Since we have immediately a field of extremales and the excess function trivializes, there is no problem about having the desired minima for our problem.) Resubstituting, we have the differential equation $(z = \sigma(y(x))/y'(x))$:

$$\frac{y'(x)}{\sigma(y(x))} = \alpha \cdot p^{1/3}(x),$$

yielding the general solutions as already described in 2.4.

B Appendix B: The Reduction of MSE by a Factor 4 on Splitting

Start (for Poisson noise case, the others work analogously) with a pleistochrome for unitary coding

$$y(x) = H^2(x), \; H'(x) = \alpha p^{1/3}(x), \; \alpha > 0.$$

Let $[a, b]$ be the range of x, $H(a) = 0$, $H(b)$ is then a fixed value > 0, $H^2(b)$ intended to be the maximal attainable firing rate. We have $MSE(y) = \frac{1}{4\alpha^2} \int_a^b p^{1/3}(x)\, dx$. Now, take the point x_0 of splitting so that $H(x_0) = \frac{1}{2}H(b)$. We are looking for two arcs, one falling on $[a, x_0]$, the other rising on $[x_0, b]$ (remember the choice to be decided only by the mean firing rate criterion, so this is just to be specific here) so that each be a squared antiderivative of some $\beta \cdot p^{1/3}$ and at each range from 0 to $H^2(b)$. The simple solution to this is $\bar{y}(x) = (2(H(x) - H(x_0)))^2$, on $[a, b]$, the corresponding primitives being $2(H(x_0) - H(x))$ on $[a, x_0]$, $2(H(x) - H(x_0))$ on $[x_0, b]$. Thus $\bar{y}(x)$ is just one of the (possibly asymmetric) U-shaped extremals discarded in 2.2. Now, since the antiderivative involved in $\bar{y}(x)$ has a factor 2 compared to that of $y(x)$, we have divided the MSE by a factor 4 according to the formula for MSE given in 2.2. Thus $MSE(\bar{y}) = \frac{1}{4}MSE(y)$. We just chose x_0 so that $H(x_0) = \frac{1}{2}H(b)$. This choice is indeed optimal:

For general x_0, $a < x_0 < b$, we have $H(x_0) = \gamma H(b)$, $0 < \gamma < 1$, with the antiderivative H mentioned above.

To adjust the ranges of the two arcs, we have a factor $\frac{1}{1-\gamma}$ for the primitive above . x_0, $\frac{1}{\gamma}$ beneath x_0. So we get

$$
\begin{aligned}
MSE(\bar{y}(x)) &= \gamma^2 \cdot c \cdot (H(x_0) - H(a)) + (1 - \gamma)^2 \cdot c \cdot (H(b) - H(x_0)) \\
&= c \cdot H(b) \cdot (\gamma^3 + (1 - \gamma)^3),
\end{aligned}
$$

c a positive constant. Now $\gamma^3 + (1 - \gamma)^3$ clearly attains its minimum with $\gamma = \frac{1}{2}$.

C   Appendix C: The Reduction of the Problem to only one Input Variable

We consider just two input variables which show how to work and need less scribbling.

Given two input variables with joint distribution $p(x_1, x_2)$ and independent noise variables n_1, n_2 with their conditional distributions $q_1(n_1 \mid y(x_1))$ and $q_2(n_2 \mid y(x_2))$, then the joint distribution of x_1, x_2, n_1, n_2 may (abbreviating) be written as $p(x_1, x_2) \cdot q_1 \cdot q_2$. The mean squared reconstruction error is then ($\hat{x}_i$ depending on $y_i(x_i)$ and n_i):

$$
\int \int dx_1\, dx_2\, p(x_1, x_2) \int \int dn_1\, n_2\, ((x_1 - \hat{x}_1)^2 + (x_2 - \hat{x}_2)^2)\, q_1 q_2.
$$

Now, split the integral in a sum, then we have for the first summand:

$$\int\int dx_1 dx_2\, p\,(x_1, x_2) \int dn_1 dn_2\, (x_1 - \hat{x}_1)^2\, q_1 q_2$$

$$= \int\int dx_1 dx_2\, p\,(x_1, x_2) \int dn_1\, (x_1 - \hat{x}_1)^2\, q_1 \underbrace{\int dn_2\, q_2}_{=1}$$

$$= \int\int dx_1 dx_2\, p\,(x_1, x_2) \int dn_1\, (x_1 - \hat{x}_1)^2\, q_1$$

$$= \int dx_1 \int dx_2\, p\,(x_1, x_2) \int dn_1\, (x_1 - \hat{x}_1)^2\, q_1$$

$$= \int dx_1 dn_1\, p_1\,(x_1)\, q_1\,(n_1 \mid y_1\,(x_1))\,(x_1 - \hat{x}_1)^2 \ ,$$

with $p_1 =$ marginal distribution, thus distribution of the input variable x_1 alone. (We did not even have to assume x_1, x_2 to be independent here, but we had to assume this to hold at least approximately to justify our simple coding $(y_1\,(x_1), \ldots, y_k\,(x_k))$ instead of $y_i\,(x_1, \ldots, x_k).$)

References

[1] Atick, J., 1992, Could information theory provide an ecological theory of sensory processing?
Network 3 (1992), 213–251

[2] Atick, J., Li, Z., and Redlich, A. N., Understanding retinal color coding from first principles, *Neural Computation 4*, 559-572

[3] Barlow, H. B., 1961, Possible principles underlying the transformation of sensory messages,
in: *Sensory Communication*, ed. W. A. Rosenblith, Cambridge, MA: MIT Press

[4] Buchsbaum, G., and Gottschalk, A., 1983, Trichromacy, opponent colours coding and optimum colour information transmission in the retina, *Proc. R. Soc. Lond. B220*, 89–113

[5] Laughlin, S. B., 1981, A simple coding procedure enhances a neuron's information capacity, *Zeitschrift für Naturforschung, 36c*, 910–912

[6] Parra, L. et al., 1995, Redundancy reduction with information-preserving nonlinear maps, *Network 6*, 61-72

18

Robert Schwartz

Directed Perception
(1/1996)

Rainer Mausfeld

"Wär' nicht das Auge sonnenhaft ..."
Goethes Farbenlehre: Nur eine Poesie
des Chromatischen oder Beitrag zu einer
naturwissenschaftlichen Psychologie?
(2/1996)

Heiko Hecht
Mary K. Kaiser
Martin S. Banks

Gravitational Acceleration as a Cue for
Absolute Size and Distance?
(3/1996)

Margaret Atherton

How to Write the History of Vision:
Understanding the Relationship between
Berkeley and Descartes
(4/1996)

Heiko Hecht

Heuristics and Invariants in Dynamic
Event Perception: Immunized Concepts
or Non-Statements?
(5/1996)

Jan J. Koenderink
Andrea J. van Doorn
Astrid M.L. Kappers
Jim Todd

The Visual Contour in Depth
(6/1996)

Jan J. Koenderink
Astrid M.L. Kappers
Frank E. Pollick
Mitsuo Kawato

Correspondence in Pictorial Space
(7/1996)

Jan J. Koenderink
Andrea J. van Doorn
Chris Christou
Joseph S. Lappin

Perturbation Study of Shading in
Pictures
(8/1996)

Jan J. Koenderink
Andrea van Doorn

How to Probe Different Aspects of
Surface Relief
(9/1996)

Jan J. Koenderink
Andrea van Doorn

The Generic Bilinear Calibration-
Estimation Problem
(10/1996)

Mark Kaplan

The Very Idea of Perception as a Process
of Unconsious Probabilistic Inference
(11/1996)

Preprintliste der Forschungsguppe *Perception and the Role of Evolutionary Internalized Regularities of the Physical World* am Zentrum für interdisziplinäre Forschung der Universität Bielefeld (1995/96), Wellenberg 1, D-33615 Bielefeld, BRD

Donald D. Hoffman Carol M. Cicerone	Dynamic color spreading and the opercular problem: A computational theory (12/1996)
Hans Irtel	Das Fechnersche Paradoxon: Folge einer Invarianzleistung der Wahrnehmung (13/1996)
Alan Nelson	Two Concepts of Idealization in Economics (14/1996)
Martin Eimer	Perception and Motor Activation Without Conscious Awareness (15/1996)
Jan Koenderink Astrid Kappers	Color Space (16/1996)
Larry Arend David Arend	Influence of Background Reflectance on Target Lightness (17/1996)
Sheng He Donald I.A. MacLeod	Contrast-Modulation Flicker: Dynamics and Spatial Resolution of the Light Adaptation Process (18/1996)
Margaret Atherton	Mr Abbot and Prof Fraser: A nineteenth century debate about Berkeley's theory of vision (19/1996)
Brian P. McLaughlin	Perception: Epistemic issues (20/1996)
Andrea J. van Doorn Alexander Toet	The Conspicuity of Targets Revealed by Relative Movement. I. Effects of target velocity, shape, and eccentricity (21/1996)
Andrea J. van Doorn Alexander Toet	The Conspicuity of Targets Revealed by Relative Movement. II. Effects of target orientation and relative target-background velocity (22/1996)

Preprintliste der Forschungsgruppe *Perception and the Role of Evolutionary Internalized Regularities of the Physical World* am Zentrum für interdisziplinäre Forschung der Universität Bielefeld (1995/96), Wellenberg 1, D-33615 Bielefeld

Andrea J. van Doorn Alexander Toet	The Conspicuity of Targets Revealed by Relative Movement. III. Drift-Balanced Versus Fourier Motion (23/1996)
Dejan Todorović	Is Kinematic Geometry an Internalized Regularity? (24/96)
Dejan Todorović	A Gem From the Past: Pleikart Stumpf's (1911) Anticipation of the Aperture Problem, Reichardt Detectors, and Perceived Motion Loss at Equiluminance (25/96)
Dejan Todorović	Lightness and Junctions (26/96)
Alan Nelson	The Falsity in Sensory Ideas: Descartes and Arnauld (27/96)
Don McLeod Tassilo von der Twer	Optimal Nonlinear Codes (28/96)