

Do Cortical Maps Adapt to Optimize Information Density?

M D Plumbley
Department of Electronic Engineering
King's College London
Strand, London WC2R 2LS, UK
Mark.Plumbley@kcl.ac.uk

Technical Report 119/SCS/98

14 October 1998

Abstract

Topographic maps are found in many biological and artificial neural systems. In biological systems, some parts of these can form a significantly expanded representation of their sensory input, such as the representation of the fovea in the visual cortex. We propose that a cortical feature map should be organized to optimize the efficiency of information transmission through it. This leads to a principle of Uniform Cortical Information Density across the map as the desired optimum. An expanded representation in the cortex for a particular sensory area (i.e. a high *magnification factor*) means that more information density is concentrated in that sensory area, leading to finer discrimination thresholds. Improvement may ultimately be limited by the construction of the sensors themselves. This approach gives a good fit to threshold vs. cortical area data of Recanzone et. al. [25, 26] on owl monkeys trained on a tactile frequency discrimination task.

1 Introduction

1.1 Topographic cortical maps

Topographic maps are a well-known feature of sensory systems in biological organisms. These maps are organized so that nearby parts of the sensory world which are represented or processed in physically nearby locations in the cortex. In vision, for example, the mammalian primary visual cortex is organized so that cortical neurons close to each other in this cortical map respond to stimuli in neighbouring parts of the visual scene. Similarly, nearby parts of the auditory cortex tend to respond to nearby frequencies; and nearby parts of the somatosensory cortex tend to respond to skin sensors in nearby parts of the skin of the animal.

One slightly strange feature of these topological maps, however, is that they can be very non-uniform in the area given over to particular parts of the sensory input. The mammalian visual cortex has a very large part given over to representation of the fovea [28]. The auditory cortex of the mustached bat has a large expanded representation around 61Hz, corresponding to the most intense (second) harmonic of its emitted echo location pulse [31]. For the cortex responsible for the sense of touch, it is well known that much more cortical area is given over to the lips and

fingers, for example, than to the back [16]. The ratio of cortical area to the area of sensory input which it represents is often known as the *magnification factor*.

Recent evidence suggests that these cortical maps can be adapted on the basis of experience, even in adult animals [8]. For example, Jenkins et. al. [14] trained adult owl monkeys to maintain digit contact with a rotating disk. They found that the receptive field sizes of the neurons in the cortex corresponding to the stimulated area of the digits became smaller (i.e. finer resolution at the skin), while the cortical area devoted to the stimulated parts of these digits increased. Thus the magnification factor in the somatosensory cortex was able to be increased through training.

This difference in cortical area of representation between different parts of the same sensory modality appears in different animals and in different sensory modalities. It therefore seems reasonable to ask the following questions:

1. Why is this non-uniformity of cortical area of representation useful to an organism?
2. What is the target for adaptation of the size of a cortical area?

In this paper we will argue that this non-uniformity in cortical area allows *efficient* processing of information in sensory modalities which have a requirement for more accurate information in some areas than in others. This will lead us to propose the principle of *Uniform Cortical Information Density*: that in order to maintain the most efficient processing of sensory information, topographic cortical feature maps adapt to maintain uniform density of Shannon information across the map. This means in turn that the magnification factor between the sensory input and the cortical map will be proportional to the input information density required. Sensory areas with higher information density will therefore have lower behavioural thresholds than those with lower information density.

Finally, we will support this argument by analyzing data from experiments reported by Recanzone et. al. [25, 26], and show that their measurements of cortical area and corresponding behavioural threshold are consistent with this principle of Uniform Cortical Information Density.

1.2 Models of cortical map formation

Many models of cortical map formation have been constructed, particularly models of the formation of orientation and ocular dominance structures in the visual cortex (see Swindale [32] for a review). Many of these are based on Hebbian learning [11], with additional lateral connections or influence which allow the topographic arrangement of the map to be formed.

In one of the early models, Willshaw and von der Malsburg [35] showed that aligned topographic maps can be formed from correlated inputs, such as clusters of 2 or 4 active input cells. The postsynaptic cells have a threshold-linear activation function, with a mexican-hat-type lateral connectivity between the cells. Takeuchi and Amari [1, 33] modified the Willshaw-von-der-Malsburg model to use a continuous neural field of threshold neurons rather than model individual neurons. They showed analytically that either a smooth topographic organization emerges, or a columnar structure will form, depending on the relative sizes of the afferent spreading and the receptive fields. The operation of their map was characterized by ‘bubbles’ of activity covering a small area within an otherwise quiet map.

Zhang [36] extends the Takeuchi-Amari neural field theory to the more general case of sigmoidal, rather than threshold, activation functions. He showed analytically that the cortical

magnification factor should be inversely proportional to the receptive field size at each point on the map. However, Petersen [21] suggests that Zhang’s uniform stimulus probability assumption means that the model cannot be used easily where a change in receptive field size is caused by non-uniform probability density. Petersen and Taylor [22] use increased probability density in the Takeuchi-Amari framework to model the training stimulus in the owl monkey experiments of Jenkins et. al. [14]. In their model, using a different derivation, a higher input probability density causes the magnification factor to increase, and the receptive field sizes to decrease.

In the well-known Kohonen self-organizing map [15], the bubbles of activity in the Amari model are replaced by a neighbourhood function operating over a two-dimensional grid of input neurons. During training, the network chooses a ‘winner’ neuron, which has the closest representation to the input. All output neurons within a the neighbourhood of the winner are adapted to be closer to the current input. In the resulting map, the input space is divided up into a number of non-overlapping regions, each owned by a particular output neuron: this is known as a *Voronoi tessellation*. An input vector within an output node’s home region will cause that neuron to fire. The resulting map exhibits an expanded representation for areas of input space which have a higher probability density, although not strictly proportional to the input pdf.

Rather than starting with a Hebb-like learning algorithm, some approaches to cortical map formation start from a target principle, such as minimization of wire length [10] or accurate modelling of the probability density of the input data space [7]. The principle we will be using here is that of efficient processing of information.

1.3 Information theoretic approaches to sensory organization

A number of authors have suggested that Shannon’s information theory [29] can help us to understand the operation of sensory systems (see e.g. [9]). Barlow [6], an early proponent of this approach, suggested that the early stages of sensory processing attempt to reduce *redundancy* in the stimulus: that is, the difference between the information content of the stimulus and the information capacity of the channels used to communicate this information. One simple way to reduce this redundancy is to ensure that channels are uncorrelated from one another, leading to the suggestion that networks with lateral inhibitory connections are trying to decorrelate their outputs.

Linsker [18] proposed that a sensory system should maximize the information transmitted through it, subject to certain constraints: this is known as his *Infomax* principle. He showed that, with certain constraints on the weights and noise on the network inputs to make the information finite, Hebbian learning algorithms that find the principal components of their input will maximize the transmitted information.

Atick and Redlich [4] take a similar information-theoretic approach, but with a slightly different emphasis which is closer to the redundancy reduction approach of Barlow. They focus on lowering the output channel capacity while maintaining the minimum information needed by the organism. In early visual processing, they believe that “the primary evolutionary pressure has been to reduce output channel capacity”, consistent with the organism keeping the complexity and cost of its sensory processing equipment as low as possible, while processing the information required. Their prediction of the ideal response of the Human Visual System based on this approach is very close to that observed psychophysically [2], and Atick, Li and Redlich [3] use this principle to predict the significant difference between neural representations of colour coding in the retinas of fish and

mammals.

Using a spiking rather than a graded neuron model, Levy and Baxter [17] considered the issue of efficiency of information transmission in a spiking neural system. They found that the optimum ratio of information capacity against energy depended on the ratio r of energy cost for the cell to emit a spike, against that for no spike. Using values of r suggested by measured glucose utilization and oxygen consumption, they calculated optimal firing frequencies close to those found in the rat subiculum (in the limbic system). They believe that the brain may have evolved to be energy efficient in its handling of information, rather than simply maximizing information capacity.

1.4 Information-theoretic cortical map models

There has already been some work on deriving self-organizing maps from information theory. For example, Linsker [19] generated a self-organizing neural map by optimizing the mutual information from an input vector to the output of a network with a single layer of output neurons, arranged in a grid. The convolution of an input-to-output point spread function and a lateral inhibition function between output neurons determines the probability of any output node firing in response to a point in the input space. For the case where the extent of the lateral inhibition dominates that of the input point spread function, the resulting input-to-node magnification factor is proportional to the input probability density, leading to uniform firing probability per output node.

Recently, Van Hulle [13] and Holthausen & Breidbach [12] both introduced the concept of constructing a cortical map through the use of node weights to define the corners of the tessellation regions, rather than as the centres of the Voronoi tessellation as used in the Kohonen map. Various information theoretic optimization principles are considered to adapt these maps. In the case of Van Hulle's algorithm, the network aims to maximize the output entropy, leading to the probability that each region captures an input vector being equal. Consequently, the resulting input-to-map magnification factor is also proportional to the input probability density.

Many of the cortical map models we have described here and earlier in this paper share one common feature: only one output node is active at any one time. Even the Amari neural field model [1], which has the potential for many active units at once, is often analyzed producing a single activity bubble centred on a single point in the map. However, this is probably not very realistic for a biological sensory system. For a given sensory input, we do not find only one neuron firing (or only one small active region), and we would not expect this to be a particularly efficient use of the representation ability of the cortex. In the rest of this paper, we will therefore start from the opposite perspective and develop a simple map model in which all output neurons can be active at the same time. We will not develop a learning algorithm for this map here, but we will consider how information theory can tell us how the map should be organized for most efficient representation of information.

2 Information density

2.1 Decorrelation and efficient information transfer

In a previous paper [24], we showed that maximization of information through a noisy channel can lead to a requirement for equal variance, decorrelated outputs from a network. We shall briefly review that result here, before we extend this concept to cortical maps.

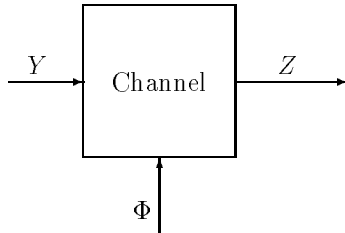


Figure 1: Signal Y is passed through a noisy channel, arriving with added noise as $Z = Y + \Phi$.

Let us represent the n outputs of a neural network as an n -dimensional random vector Y . This is transmitted via a channel (Fig. 1) which has added noise Φ , giving a received signal at the other end of the channel of

$$Z = Y + \Phi. \quad (1)$$

Suppose that the signal Y and noise Φ are gaussian with covariance matrices $\mathbf{C}_Y = E(YY^T)$ and $\mathbf{C}_\Phi = E(\Phi\Phi^T)$ respectively, and that Φ is small. Then the information transmitted from Y to Z is

$$I(Z, Y) = \frac{1}{2} \log \det \mathbf{C}_Z - \frac{1}{2} \log \det \mathbf{C}_\Phi \quad (2)$$

$$\approx \frac{1}{2} \log \det \mathbf{C}_Y - \frac{1}{2} \log \det \mathbf{C}_\Phi \quad (3)$$

with a power cost to transmit the information of

$$S_T = \text{Tr}(\mathbf{C}_Y). \quad (4)$$

where $\text{Tr}(\cdot)$ is the matrix trace operator. Now, we wish to transmit our information as efficiently as possible, so we wish to maximize $I(Z, X)$ for a given value of S_T . Using the technique of Lagrange multipliers, we therefore attempt to maximize the function

$$J = I(Z, Y) - \frac{1}{2} \lambda S_T. \quad (5)$$

where λ is our Lagrange multiplier. We would get an equivalent expression were we to minimize the power cost S_T for fixed information $I(Z, X)$, and this is also equivalent to minimizing redundancy [23].

Taking the derivative of eqn. (5) with respect to \mathbf{C}_Y and equating to zero leads to the condition [24]

$$\mathbf{C}_Y = (1/\lambda) \mathbf{I}_n \quad (6)$$

where \mathbf{I}_n is the identity matrix. Thus for most efficient transmission of information through the channel, we should ensure that the network outputs are uncorrelated and have equal variance $1/\lambda$.

If Y is non-gaussian, then the requirement that we have uncorrelated, equal variance outputs is not sufficient to maximize the transmitted information as such. However, we can say that this condition maximizes the *information capacity*

$$C(Z, Y) = \max_Y I(Z, Y) \quad (7)$$

for any given power limit S_T , and that this capacity will be achieved if Y is gaussian.

For simplicity in the rest of this paper, we will assume that the signal Y is, in fact, gaussian. If this is not the case, we can simply repeat the analysis substituting ‘information capacity’ for ‘information’. Note that we must be more careful with the noise term Φ . Were this not to be gaussian, the disruption to the signal may be *reduced*, and it may be possible that *more* information is transmitted than we were expected. If we wanted to allow non-gaussian noise, we would have to ensure that the effect of the noise was not underestimated by the equivalent gaussian noise with covariance matrix \mathbf{C}_Φ .

2.2 Information density in a discrete map

Consider the distribution of information capacity in the n neurons in the previous section. At the optimum point, the output variance σ_Y^2 and noise variance σ_Φ^2 of all neurons is equal, and hence the information

$$I_i = \frac{1}{2} \log(1 + \sigma_Y^2 / \sigma_\Phi^2) \quad (8)$$

is equal for each neuron i , where $\sigma_Y^2 = S_T/n$. Since the maximum information is achieved for a gaussian signal with uncorrelated (and hence independent) outputs, the total information $I(Z, Y)$ transmitted across the entire map is simply the sum of the information capacity of the individual neurons, i.e. $I(Z, Y) = nI_i$.

At present, these neurons have no particular order. Now suppose we arrange them into a two-dimensional map. Let us assume that the neurons are the same size, and hence packed into the map with uniform density h neurons per unit area. Therefore the map will have uniform information density of

$$I' = \frac{1}{2} h \log(1 + \sigma_Y / \sigma_\Phi) \quad (9)$$

$$= \frac{1}{2} h \log(1 + S_T / N_T) \quad (10)$$

where

$$N_T = \text{Tr}(\mathbf{C}_\Phi) \quad (11)$$

is the total noise power. If the logarithm in eqn. (10) is to the base 2, I' will be expressed in bits per unit area: if it is a natural logarithm, I' will be in nats per unit area. Unless stated otherwise, will use natural logarithms in this paper for mathematical convenience.

This expression leads us to propose the following organizational principle for cortical maps: that the most efficient processing of sensory information in a topographic cortical map with uniform density of neurons will be achieved when the information density across the map is uniform. We shall refer to this as the principle of *Uniform Cortical Information Density* or just *Uniform Information* for short.

2.3 Linear filtering

A useful analogy to this map is the consideration of the filtering of a signal in the frequency domain [30]. Consider a normally distributed signal $Y(v)$ with mean zero and variance $S(v) = E(Y(v)^2)$ at v . In the original frequency domain approach, v is a frequency f : later in this paper we

will allow it to be a more general location in an information-bearing medium such as a cortical surface. Suppose that the signal $Y(v)$ is subject to additive noise of variance $N(v)$, and total power (variance) limit $S_T = \int S(v)$.

Given a noise profile $N(v)$, the total information capacity of this system is maximized when $S(v) + N(v)$ is a constant [30]. For example, if $N(v) = N$ is constant with v (so-called *white* noise), then we should have $S(v) = S$. Thus the information capacity density $0.5 \log(1 + S(v)/N(v))$ should also be uniform.

Suppose we are considering linear filters, where $v = f$ represents the frequency of the signal component $Y(f)$. The function of a linear filter is to multiply frequency components of an input signal $X(f)$ by a gain factor $G(f)$ to give the resulting output

$$Y(f) = G(f)X(f). \quad (12)$$

If $S_X(f) = E(X(f)^2)$ is the variance of the input signal, and $S_G(f) = G(f)^2$ is the power gain of the filter, then the optimum condition will be achieved if

$$S_G(f) \propto 1/S_X(f). \quad (13)$$

This type of filter is called a *whitening filter*. Normally there is a bandwidth limit $f \leq W$, and dependencies between different frequency components are ignored (if there were dependencies, this would reduce the actual information transmitted to less than the information capacity available).

This arrangement achieves our most efficient transmission of information through the noisy channel (or cortex) by weighting the frequency components of the input signal until they have uniform variance, before transmission through the noisy channel. Thus this achieves our goal, but at a price: the density of information at all frequencies f is the same, so the information that can be transmitted about the signal content within any frequency range Δf is identical no matter what frequency that range is about. Thus in the case of a linear filter, the uniform information density requirement has forced the sensory input domain to also have uniform information density.

Consider the implications of this arrangement for the bat auditory system, for example. For the bat to fly and hunt effectively, much more accurate information is needed about sounds near to the frequencies used by the echo location pulses than for other frequencies. Rather than the simple linear filtering approach, which would force the same information density on all frequencies, a more flexible approach is needed.

2.4 Manipulating information density

Suppose now that the input to our system is a random function $X(u)$, which is defined over a different continuous domain u to the domain v over which the output $Y(v)$ is defined. Suppose also that we have a continuous map from an input range u to an output range $v(u)$. To make what follows meaningful, we will further suppose that the output $Y(v)$ is determined from $X(u)$ ‘locally’ to $v(u)$, i.e. changes to $X(u)$ will cause the most significant effects in $Y(v)$ ‘close’ to $v(u)$. Thus information that is concentrated around u in a sensory input X will be concentrated around the point $v = v(u)$ in the cortical map Y .

Let us allow the local scaling dv/du in the forward direction to vary: this is the *magnification factor* for the mapping from $X(u)$ to $Y(v)$. Let us denote the information density in the input as

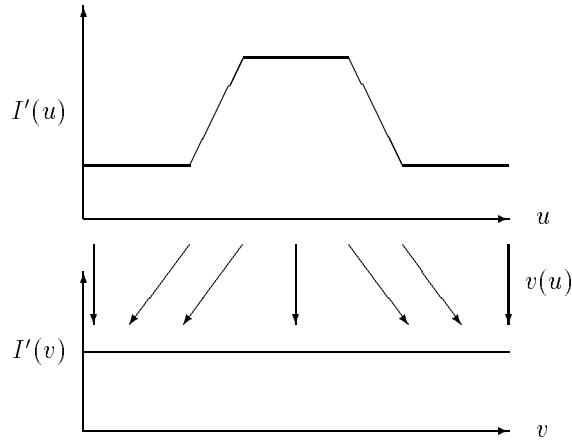


Figure 2: Magnification of input range over some regions of input u , allowing a non-uniform information density $I(u)$ over u , while we have a uniform map information density $I(v)$ over v .

$I'(u)$ and in the output as $I'(v)$. The information transmitted through a region dv about a point v in the cortical map is the same as that transmitted through the corresponding region du about a point u in the sensory input. Therefore we have

$$I'(u) du = I'(v) dv \quad (14)$$

i.e.

$$I'(u)/I'(v) = dv/du. \quad (15)$$

For optimum efficiency, we require the information density $I'(v)$ in the map v to be uniform. Under this condition, the information density in the input is given by

$$I'(u) = k(dv/du) \quad (16)$$

where $k = I'(v)$ is the map information density. Thus we have a corollary of the Uniform Cortical Information Density principle: that the input-to-map magnification factor dv/du at a point in the input space is proportional to the information density $I'(u)$ required at that point (fig. 2).

2.5 Information and behavioural thresholds

What does it mean when we say that we have increased the information density around a point? To answer this, suppose that we have a small region in our sensory input around u of size Δu with information density $I'(u)$. Thus the total information available in that region would be

$$I_{\Delta u} = I'(u)\Delta u \quad (17)$$

Now, let us suppose that all of this information capacity in the region Δu is being used to send information about a single normally distributed random variable with variance S (this would probably require a very complex encoding in practice, but the implementation will not concern us here). We will further assume that we can model the behavioural threshold θ for this signal S as a gaussian noise term with variance N such that $\theta = \sqrt{N}$. The information available after the

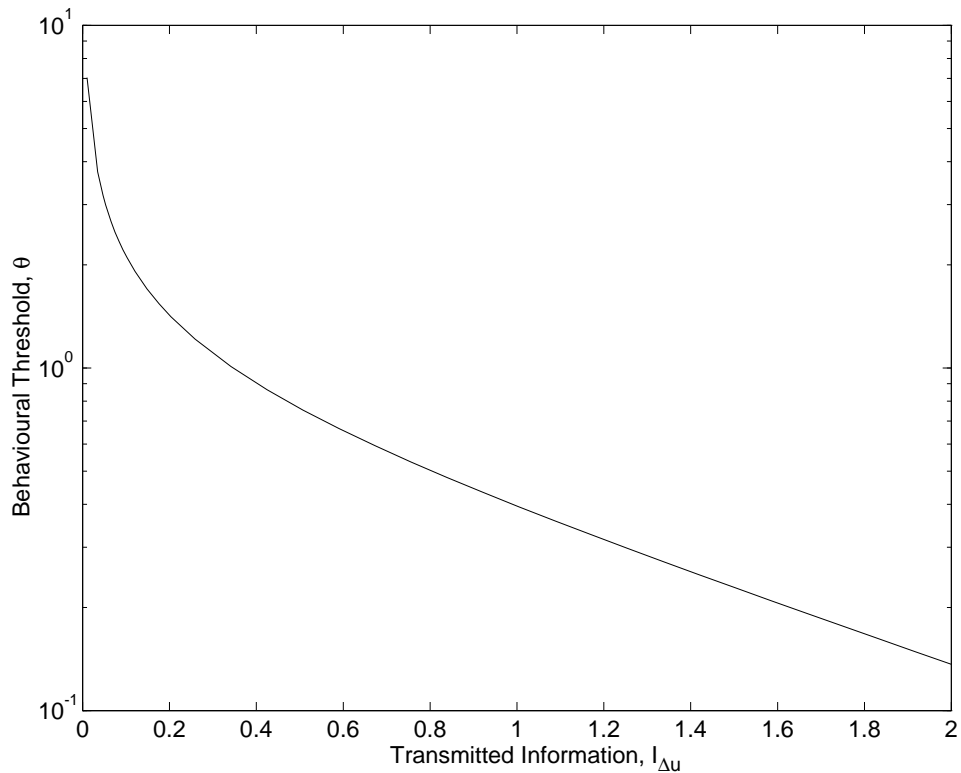


Figure 3: Plot of minimum behavioural threshold (noise standard deviation) against information density with $S = 1$.

cortical map can be no more than the information available around the region Δu , therefore N must be such that the following equation holds:

$$I_{\Delta u} = I'(u)\Delta u \geq \frac{1}{2} \log(1 + S/N) \quad (18)$$

or

$$\theta = \sqrt{N} \geq \sqrt{S(\exp(2I_{\Delta u}) - 1)^{-1}} \quad (19)$$

which is plotted in Fig. 3 for $S = 1$. Thus a higher magnification factor leads to higher information density, a higher amount of information through the region Δu , and consequently a lower behavioural threshold.

From this viewpoint, the purpose of different magnification factors in sensory cortex is therefore that these allow different behavioural thresholds, but while keeping the overall system as efficient as possible.

2.6 Incorporating sensor noise

So far we have ignored any noise in the sensors themselves. Without such noise, Fig. 3 suggests that we can reduce the behavioural threshold as much as we like by allocating it sufficient cortical

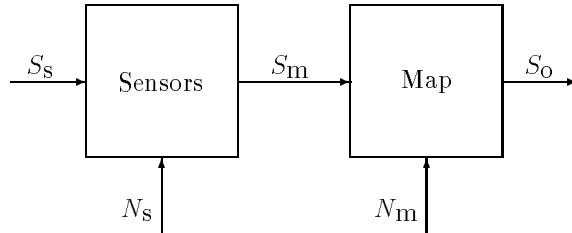


Figure 4: In the scalar linear mixing model, we assume that independent noise in the skin surface receptors (‘Sensors’) and cortical map (‘Map’) is added to give the effective noise on the signal. For convenience, the gain of the skin sensor and cortical map systems is scaled to unity.

area Δv and hence increasing the information density at the sensors. However, if the sensors themselves are noisy, they will force a lower limit to this behavioural threshold.

Let us model the sensor-cortex sensory system as an information transmission system with two additive noise sources at two locations: at the skin surface receptors and after the neurons in the cortical map. According to our uniform information density hypothesis, the cortical map representation should expand where increased information capacity is required so that the area of any cortical map sector is proportional to the capacity of that sector. However, we will assume that the corresponding skin surface receptors are fixed and cannot adapt in the same way. We shall therefore assume that the information capacity of the skin surface receptors themselves remains constant, and that this information capacity will consequently be an upper limit to the information capacity of the entire sensor-cortex information transmission system.

We shall again represent signal and noise sources in this system as simple normally distributed (i.e. gaussian) random variables, where the noise sources are additive and independent of the signal (and each other). Without loss of generality, signal gain will be normalized to unity throughout.

We will consider a simple linear mixing model (Fig. 4). Here we model the stimulus signal as a scalar gaussian (normally distributed) random variable with variance S_s . At the skin surface receptors, this is corrupted by independent additive gaussian noise of variance N_s , giving a resulting gaussian signal-plus-noise of variance

$$S_m = S_s + N_s \quad (20)$$

which represents the output from the skin receptors that is transmitted to the cortical map segment.

Suppose that the cortical map segment operates at an information density $I'(v) = k$ and has area A , giving a total information capacity for the map segment of $I_m = kA$. In this model, this limited information capacity I_m is modelled by an effective additive gaussian noise of variance N_m introduced by the cortical map itself, such that the total information capacity of the map,

$$I_m = \frac{1}{2} \log \left(1 + \frac{S_m}{N_m} \right) \quad (21)$$

is the information capacity of a scalar gaussian channel with signal variance S_m and noise variance N_m [29]. Note that the ‘signal’ S_m does in fact contain some noise N_s due to the skin receptors, but we suppose that the cortical map is not able to separate this added noise from the initial input

signal S_s . The final signal output of the skin-cortex system is then

$$S_o = S_m + N_m = S_s + N_s + N_m. \quad (22)$$

Re-arranging eqn. (21) for N_m and substituting in eqn. (20), we get

$$N_m = (S_s + N_s)(\exp(2kA) - 1)^{-1} \quad (23)$$

resulting in an overall effective gaussian noise for the skin-cortex system of variance

$$N_{\text{eff}} = N_s + N_m = N_s + (S_s + N_s)(\exp(2kA) - 1)^{-1}. \quad (24)$$

We take an experimentally-recorded behavioural threshold level θ to correspond to one standard deviation of this overall effective gaussian noise [25]. Therefore, as the cortical map area A devoted to processing a particular input signal changes, we would expect the behavioural threshold θ also to change, and be determined from A according to the following formula:

$$\theta = \sqrt{N_{\text{eff}}} = \sqrt{N_s + (S_s + N_s)(\exp(2kA) - 1)^{-1}} \quad (25)$$

which is plotted in Fig. 5 for various levels of N_s .

Thus if we know the sensor noise N_s , the signal variance S_s , and the optimal cortical information density $I'(v) = k$ for a particular sensory system, we should be able to use eqn. (25) to predict how the behavioural threshold of a sensory system will vary with area. In the next section, we will use this to investigate the results of Recanzone and colleagues on expansion of the somatosensory cortex in adult own monkeys.

3 Comparison with experiment

3.1 Expansion in the somatosensory cortex

In a series of experiments, Recanzone et. al. [25] trained owl monkeys on a tactile frequency discrimination task. The animals were trained to detect frequency differences above a 20Hz baseline of a vibrating stimulus applied to a small part of one finger. Over a number of weeks, discrimination thresholds on the trained finger decreased from around 6–8Hz above the baseline to 2–3Hz above the baseline. Following training, the cortical maps of the trained animals showed a larger representation of the stimulated skin in the primary somatosensory cortex (S1 area 3b) when compared to control skin sites, suggesting that the cortical representation area had increased due to training.

They found that a large area of cortical representation for the stimulated parts of digits tended to go with a lower behavioural threshold, although a simple straight line fit did not demonstrate significant dependence to the 99% level that they were using [26].

The correlation they observed would be expected from our hypothesis above, that an increase in cortical area leads to a increase in information capacity, since an increase in information capacity would imply a decrease in the effective noise level on the perceptual signal due to processing in the cortical map. Using our uniform information density principle we shall therefore try to show that there is in fact a significant dependence between these two parameters, and that this dependence is consistent with our principle.

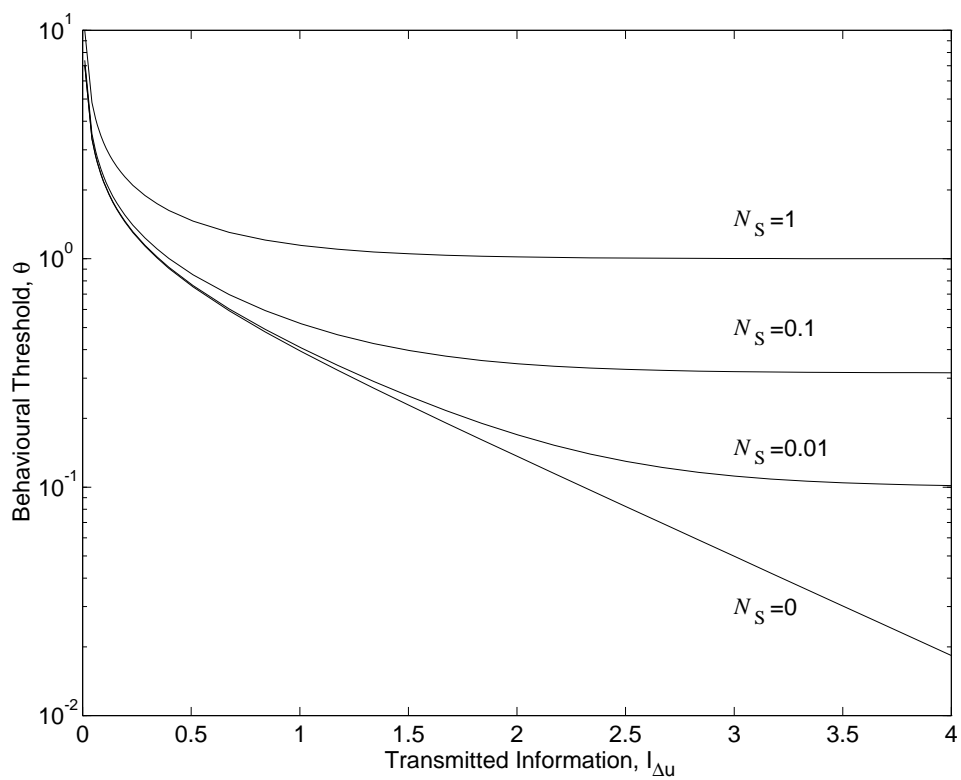


Figure 5: Minimum behavioural threshold plotted against information $I_{\Delta u}$ for noisy sensors at various levels of sensor noise N_S . For this plot we have $S_S = 1$.

3.2 Measurements

Discrimination thresholds in Hz were reported by Recanzone et. al. [26, Table 1] for a trained digit (TD) and an adjacent digit (AD) on each of five animals (E1, . . . , E5), giving 10 measurements in total. For these animals, the cortical area of representation A was also reported for the stimulated parts of both the trained and adjacent digits [26, Fig. 12]. Two further animals (E6 and E7) were also trained on the same task (two digits, E6-D3 and E6-D4, on animal E6) but without measurements of cortical area [25, Table 1].

In [27] they argued that the temporal response time of cortical neurons is a critical factor for this task. Therefore in this paper, we will use stimulation cycle period discrimination threshold (in milliseconds) rather than a frequency threshold (in Hz) as the behavioral threshold θ . The original frequency thresholds above a 20 Hz baseline become cycle period thresholds below a 50 msec baseline. The animal training sessions used periods in the range 33–50 msec (frequencies in the range 20–30Hz).

3.3 Estimation of parameters

In order to generate a curve of behavioural threshold against cortical area, we require 3 parameters to be estimated: the sensor signal variance S_s , the sensor noise variance N_s , and the information density k .

While the training sessions used a limited range of cycle periods (33–50 msec) we will assume that the sensors themselves have a wider range than this. We will assume that the signal range of the sensors corresponds to a normal distribution centred at 50 msec with 3 standard deviations out at 0 msec and 100 msec. Thus we will use $S_s^{1/2} = 50/3 \approx 16.67$ msec.

To estimate the sensor noise variance N_s , we will assume that after significant training the map area has expanded sufficiently such that the sensor noise dominates, so that the behavioural thresholds at the final training session mainly represent that due to the sensor noise. To ensure we do not use data points more than once, we will use the average of the thresholds reported without corresponding cortical area measurements. For E6-D3 E6-D4 and E7 these were 1.62, 1.53, and 2.87 Hz respectively, i.e. 3.74, 3.55 and 6.27 msec respectively, giving an average of 4.52 msec for sensor threshold. We will use this as 1 standard deviation of the sensor noise pdf, so that $N_s^{1/2} = 4.52$ msec.

The information density k cannot be estimated in isolation, so we will use this parameter to fit the best curve to the reported data values.

3.4 Fitting the curve to the data

The curve fit was carried out using Matlab 4.0. Before fitting, the thresholds θ and area A were normalized to unit variance. The information density k was then adjusted using the Matlab Simplex search function ‘fmins’ to find the least mean squared error fit. The mean squared error was calculated parallel to the line $A/\sigma_A = \theta/\sigma_\theta$, i.e. at a constant angle of $\pi/4$ from the horizontal on a normalized plot of k against A . On each iteration the point on the curve corresponding to each data point was itself found using ‘fmin’. Both minimisation steps used the default termination tolerance of 10^{-4} . The use of a constant projection angle ($\pi/4$), rather than a more flexible total least squares fit, does give us a slightly worse mean squared error value than we would have if we

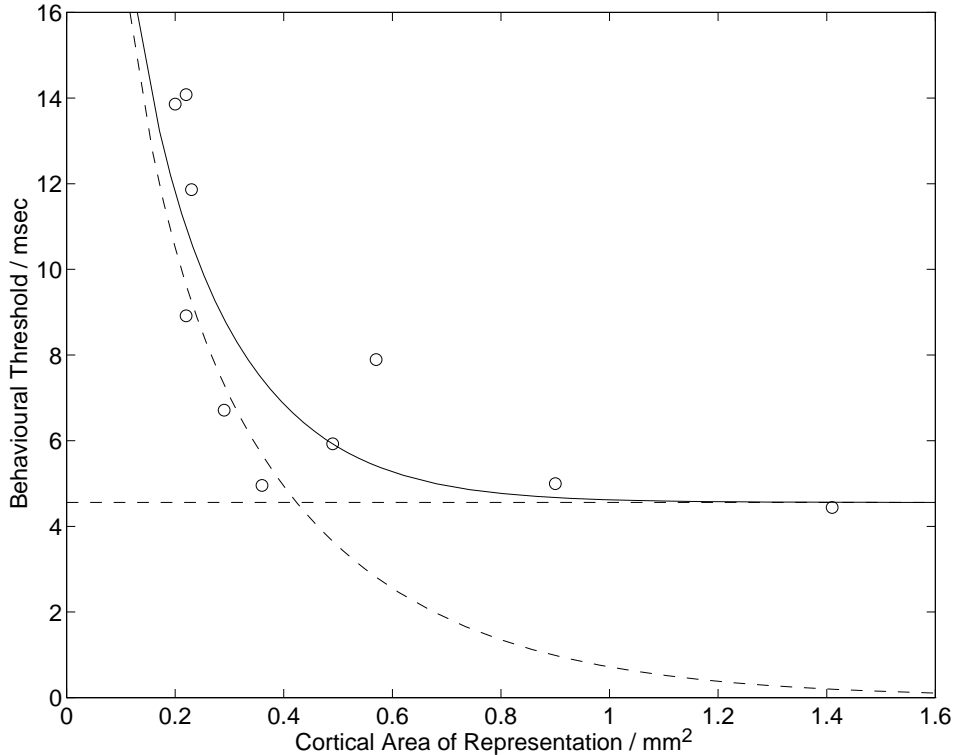


Figure 6: Best fit of the curve defined by eqn. (25) (solid curve) to the data (circles). We use $S_S = 50/3$ msec and $N_S = 4.525$ msec in eqn. (25), and the best fit (as described in the text) is achieved for $k = 4.515$ bits/mm², leaving a resulting normalized mean squared error of 0.0625. The threshold due to the sensor alone (dashed horizontal line) and the threshold due to the map alone (dashed curve) are also shown.

looked for the closest point on the curve for each data point, but it is conservative enough to allow us to perform a simple significance test (see Appendix).

Fig. 6 shows the experimental data plotted with the best fit curve, as described above. The best fit was achieved with $k = 3.130$ nats/mm² = 4.515 bits/mm², giving a normalized mean squared error (NMSE) of $E = 0.0625$. The parameters are summarized in Table 1. This fit is significant, in the sense that this curve is consistent with a dependence between A and θ sufficient to reject the hypothesis that they are independent at the 99% confidence level (see Appendix).

4 Discussion

We have shown that our principal of uniform cortical information density is consistent with the experimental observations described above. Furthermore, this gives us an estimate of for the information density in the cortex of $k \approx 4.5$ bits/mm². This seems rather on the low side for what we might expect. However, the frequency discrimination task considered here is probably

$S_S^{1/2}$	$N_S^{1/2}$	$k/\log_e(2)$	E
msec	msec	bits/mm ²	(NMSE)
16.67	4.525	4.515	0.0625

Table 1: Summary of best fit curve parameters shown in Fig. 6, together with the resulting normalized mean squared error (NMSE) E .

not one that the somatosensory cortex is designed to handle well. We might therefore expect that the the information capacity of this part of the cortex for more realistic sensory input would be significantly higher than this.

It may be noticed that we have not constructed an adaptation algorithm to generate this organization. Rather we have specified a target function that we would wish an ideal algorithm to perform. In fact, it would be perfectly possible for the area A in the map to be distributed in a very non-local manner, provided all the contributions could be identified. We therefore believe that the optimal information density approach is independent of any locality of organization in the map.

A further efficiency issue that we have not considered in this paper is that of the cost of wiring [10]. It may be that minimisation of wirelength cost could be an orthogonal principle to that of uniform information density, and would give us the locality of representation that we see in biological neural maps.

For the experimental results we considered here, the cortical map adapted to increase the information about the sensory input that was relevant to the task, but we presume that the discrimination ability was ultimately limited by the existing skin sensor arrangement. However, over generations, it would be expected that evolution could rearranged sensors if this was important for the animal's survival. As we have seen, information efficiency is normally optimized if the information is evenly spread across the available channels, if they are all of identical construction. We would therefore expect the uniform information density in the cortex to be reflected in a uniform use of sensors themselves. In other words, the cortical magnification factor should match the density of sensors, and we would expect a constant scale mapping from sensor cells to cortical cells.

There are suggestions that the mapping from retinal ganglion cells to visual cortex cells is roughly constant in scale, except in the fovea [34]. Here the ganglion cells are not as dense as the cortical magnification factor would suggest. It may be that physical limits in the retina [5] mean that it is not possible for the ganglion cells, or their corresponding receptors, to be packed more densely in the retina. Therefore if the cortical magnification factor is to be reflected in an increased information density in the fovea, this would have to be achieved another way. On this basis we would predict that to compensate, ganglion cells would have a higher mean firing rate in the fovea than in the periphery, in order to improve their signal-to-noise ratio to match the cortical information density more closely.

5 Conclusions

In this paper, we have proposed that cortical maps adapt to represent the sensory information that they need as efficiently as possible. This leads is to develop a concept of *information density* over

a cortical map. The optimum arrangement is for this information density to be uniform across the map: we refer to this as the principal of *Uniform Cortical Information Density*, or more simply *Uniform Information*.

From this viewpoint, the cortical magnification factor allows the sensory information density to be concentrated where it is required, leading to lower behavioural discrimination thresholds. For example, in the bat, having a large magnification factor for auditory frequencies around the frequency of the echo location pulse will allow more information to be extracted around that frequency than around others. If the sensors themselves have significant levels of noise, these will impose an upper limit on the information density, and thus a lower limit on any discrimination thresholds.

This model was found to give a good fit to data of Recanzone et. al. [25, 26] on owl monkeys trained on a tactile frequency discrimination task. This gave an estimate of $k \approx 4.5\text{bits}/\text{mm}^2$ for the cortical information density, although this is likely to be an underestimate of the true value due to the unusual nature of the task.

It is hoped that this Uniform Information principle will be capable of making testable predictions about the arrangement of cortical maps, and of forming a basis for the development of future learning algorithms for these maps.

Acknowledgements

I would like to thank John Shawe-Taylor and the rest of Department of Computer Science at Royal Holloway, University of London, for the constructive and enjoyable time I spend there during the summer of 1997, while some of this work was carried out.

References

- [1] S.-I. Amari. Field theory of self-organizing neural nets. *IEEE Transactions on Systems Man and Cybernetics*, SMC-13:741–748, 1983.
- [2] J. J. Atick. Could information-theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, 3:213–251, 1992.
- [3] J. J. Atick, Z. P. Li, and A. N. Redlich. Understanding retinal color coding from first principles. *Neural Computation*, 4:559–572, 1992.
- [4] J. J. Atick and A. N. Redlich. Towards a theory of early visual processing. *Neural Computation*, 2:308–320, 1990.
- [5] H. B. Barlow. Critical limiting factors in the design of the eye and visual cortex. *Proceedings of the Royal Society of London B*, 212:1–34, 1981.
- [6] H. B. Barlow. Unsupervised learning. *Neural Computation*, 1:295–311, 1989.
- [7] C. M. Bishop, M. Svensen, and C. K. I. Williams. GTM: A principled alternative to the self-organizing map. Technical Report NCRG/96/015, Neural Computing Research Group, University of Aston, UK, 1996.

- [8] A. Das. Plasticity in adult sensory cortex: A review. *Network: Computation in Neural Systems*, 8:R33–R76, 1997.
- [9] G. Deco and D. Obradovic. *An Information-Theoretic Approach to Neural Computing*. Springer, New York, 1996.
- [10] R. Durbin and G. Mitchison. A dimension reduction framework for understanding cortical maps. *Nature (London)*, 343:644–647, 1990.
- [11] D. O. Hebb. *The Organization of Behavior*. Wiley, New York, 1949.
- [12] K. Holthausen and O. Breidbach. Self-organized feature maps and information theory. *Network: Computation in Neural Systems*, 8:215–227, 1997.
- [13] M. M. Van Hulle. Topology-preserving map formation achieved with a purely local unsupervised competitive learning rule. *Neural Networks*, 10:431–446, 1997.
- [14] W. M. Jenkins, M. M. Merzenich, M. T. Ochs, T. Allard, and E. Guíc-Robles. Functional reorganization of primary somatosensory cortex in adult owl monkeys after behaviorally controlled tactile stimulation. *Journal of Neurophysiology*, 63:82–104, 1990.
- [15] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.
- [16] S. W. Kuffler, J. G. Nicholls, and A. R. Martin. *From Neuron to Brain: A Cellular Approach to the Function of the Nervous System*. Sinauer Associates Inc., Sunderland, MA, second edition, 1984.
- [17] W. B. Levy and R. A. Baxter. Energy efficient neural codes. *Neural Computation*, 8:531–543, 1996.
- [18] R. Linsker. Self-organization in a perceptual network. *IEEE Computer*, 21(3):105–117, March 1988.
- [19] R. Linsker. How to generate ordered maps by maximising the mutual information between input and output signals. *Neural Computing*, 1:402–411, 1989.
- [20] D. H. Menzel, editor. *Fundamental Formulas of Physics*, volume 1. Dover Publications, New York, 1960.
- [21] R. S. Petersen. *A Neural Field Theory Approach to Cortical Self-Organisation*. PhD thesis, Department of Mathematics, King’s College London, 1998.
- [22] R. S. Petersen and J. G. Taylor. Reorganisation of somatosensory cortex after tactile training. In D. S. Touretsky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 82–88. MIT Press, 1996.
- [23] M. D. Plumbley. On information theory and unsupervised neural networks. Technical Report CUED/F-INFENG/TR.78, Cambridge University Engineering Department, UK, 1991.
- [24] M. D. Plumbley. Efficient information transfer and anti-Hebbian neural networks. *Neural Networks*, 6:823–833, 1993.

- [25] G. H. Recanzone, W. M. Jenkins, G. T. Hradek, and M. M. Merzenich. Progressive improvement in discriminative abilities in adult owl monkeys performing a tactile frequency discrimination task. *Journal of Neurophysiology*, 67:1015–1030, 1992.
- [26] G. H. Recanzone, M. M. Merzenich, W. M. Jenkins, K. A. Grajski, and H. R. Dinse. Topographic reorganization of the hand representation in cortical area 3b of owl monkeys trained in a frequency-discrimination task. *Journal of Neurophysiology*, 67:1031–1056, 1992.
- [27] G. H. Recanzone, M. M. Merzenich, and C. E. Schreiner. Changes in the distributed temporal response properties of SI cortical neurons reflect improvements in performance on a temporally based tactile discrimination task. *Journal of Neurophysiology*, 67:1071–1091, 1992.
- [28] J. Rovamo, V. Virsu, and R. Näsänen. Cortical magnification factor predicts the photopic contrast sensitivity of peripheral vision. *Nature*, 271:54–56, 1978.
- [29] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.
- [30] C. E. Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37:10–21, 1949.
- [31] N. Suga. Cortical computational maps for auditory imaging. *Neural Networks*, 3(1):3–21, 1990.
- [32] N. V. Swindale. The development of topography in the visual cortex: A review of models. *Network: Computation in Neural Systems*, 7:161–247, 1996.
- [33] A. Takeuchi and S. Amari. Formation of topographic maps and columnar microstructures in nerve fields. *Biological Cybernetics*, 35:63–72, 1979.
- [34] V. Virsu, R. Näsänen, and K. Osmoviita. Cortical magnification and peripheral vision. *Journal of the Optical Society of America A: Optics and Image Science*, 4:1568–1578, 1987.
- [35] D. J. Willshaw and C. von der Malsburg. How patterned neural connections can be set up by self-organization. *Proceedings of the Royal Society of London, B*, 194:431–445, 1976.
- [36] J. Zhang. Dynamics and formation of self-organizing maps. *Neural Computation*, 3:54–66, 1991.

Appendix: Statistical significance

To test the significance of the curve fit, we will attempt to reject the null hypothesis that θ and A are independent. Recall that we had performed a curve fit to $n = 10$ data points with normalized mean squared error $E = 0.0625$ measured in the $A/\sigma_A = \theta/\sigma_\theta$ direction.

Suppose we had fit the normalized data with $n = 10$ data points using a straight line in the $A/\sigma_A = -\theta/\sigma_\theta$ direction, giving a normalized mean squared error $E = 0.0625$ in the $A/\sigma_A = \theta/\sigma_\theta$ direction (i.e. the value we had observed for our curve). We would therefore have an observed correlation coefficient of

$$r = - \left(1 - \frac{n}{n-1} E \right) = -0.9306. \quad (26)$$

If the null hypothesis were true, then the true correlation coefficient ρ would be zero, and the statistic

$$t = \left| \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \right| = 7.1883 \quad (27)$$

would follow the Student t -distribution for $(n-2)$ degrees of freedom (see e.g. [20]). The value of $t = 7.1883$ above is sufficient to reject this hypothesis at the 99% confidence level.

However, rather than a straight line fit, we have a curve fit, but with the same normalized mean squared error E and in the same direction. Let us therefore reduce the curve to a straight line by translating all points in a direction parallel to the $A/\sigma_A = \theta/\sigma_\theta$ direction (i.e. the $\pi/4$ direction) so that the solid curve in Fig. 6 is ‘straightened out’ and lies in the $A/\sigma_A = -\theta/\sigma_\theta$ direction. If A and θ were independent, the confidence that E measured in the $A/\sigma_A = \theta/\sigma_\theta$ direction would not be significantly small following this distortion, and hence that r would not be significantly negative, would be greater than if we had not performed this distortion. Thus the fact that E is sufficiently small to reject the null hypothesis were there to be no distortion means that we can also reject the null hypothesis for the current case.