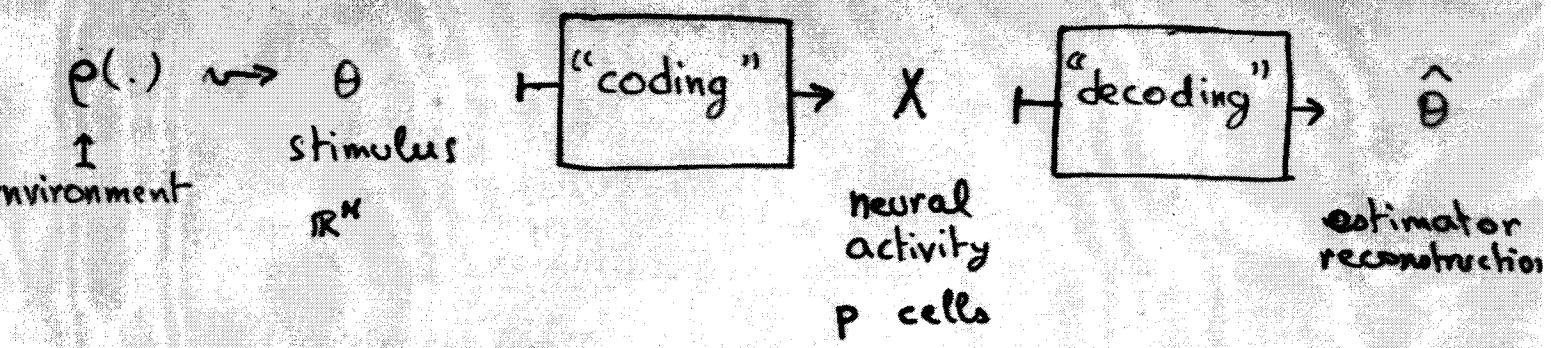
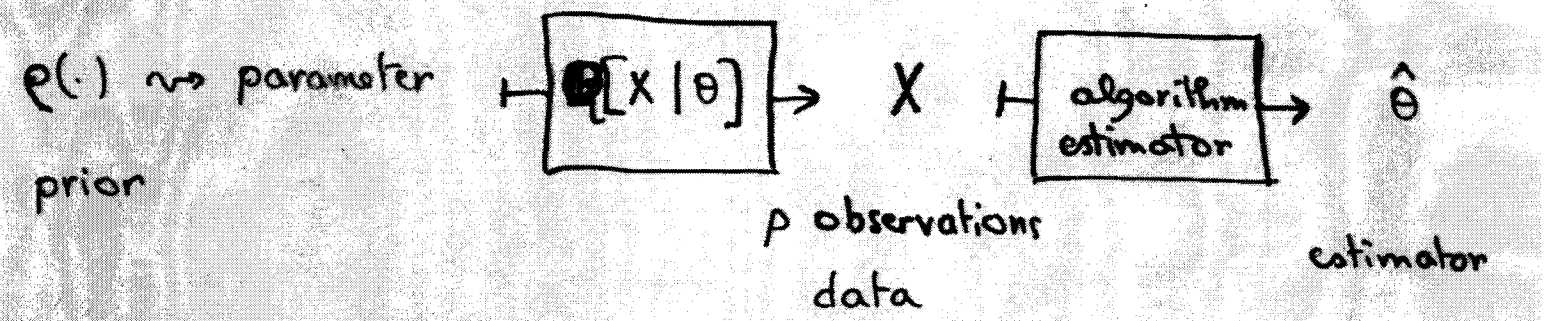
– neural coding –



– Bayesian inference –



- coding efficiency
- optimal performances
- performance of a given estimator

Mutual information (Shannon)
Fisher information
Redundancy

$I[\theta; X]$ information conveyed by X about θ

statistical dependency between X and θ

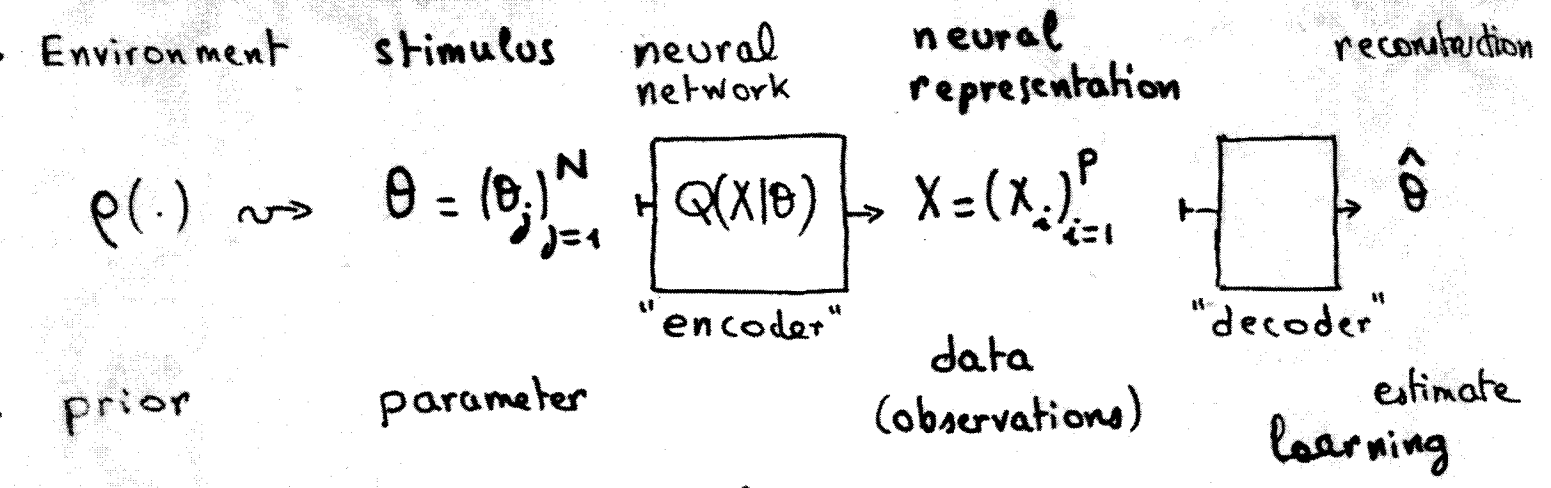
Knowledge of θ is limited : finite amount of data

independant of any specific algorithm for decoding / estimation of θ

optimal estimator = extract all this information hidden in the data

- $C = \max_{\rho} I$ (given Q) capacity $Q(X \mid \theta)$
- $I_m \equiv \max_{Q} I$ (given ρ) (infomax) adaptation to the environment
- large N and p : typical $I(\theta; X)$ (stat. mech.)
- rigorous bounds and asymptotic behaviour (p large)
- behaviour of the mutual information (N, p)
- $N = p = 1$ single cell : optimization of the transfer function
  ($\leftrightarrow$ image processing : histogram equalization)

neural coding



Bayesian learning

---

* common tool:

mutual information $I[\theta; X] = \int d^N\theta\, \rho(\theta) \int d^P X\, Q(X|\theta) \log \frac{Q(X|\theta)}{Q(X)}$

$= \int d^N\theta\, d^P X\, P(X,\theta) \log \frac{P(X,\theta)}{Q(X)\rho(\theta)}$

I = Bayes risk = cumulative relative entropy loss [Haussler-Opper 1997]

• examples:

neural coding:
- N small $\ll P$ large — population coding
- $N \ll P$ — sparse coding

learning:
- $1 \ll N \ll P$ — efficient generalization
- $N = 1 = P$ — equalization ($f' = \rho$)

• population coding: θ = angle X_i = {firing rate of cell i / or: number of spikes}

e.g.: Poisson model: $Q_i(X_i = n|\theta) = \frac{[\nu_i(\theta)t]^n}{n!} e^{-t\nu_i(\theta)}$

$\nu_i(\theta) = \nu\left(\frac{\theta - \theta_i}{\sigma_i}\right)$ "tuning curve"

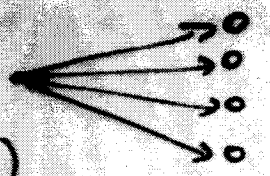
• supervised learning: $X_i = (\vec{x}^{(i)}, \tau^i)$ $\vec{x}^i \in \mathbb{R}^N$, $\tau^i = \pm 1$

$\tau^i = \mathrm{sgn}(\theta \cdot \vec{x}^i)$ θ = "teacher"

$Q(X|\theta) = \prod_{i=1}^P Q_i(X_i|\theta)$

Population Coding

$S = \theta$ (e.g. direction)



p cells p large

Simple model: Poisson processes

output cell i : $\mathbb{P}(k_i \text{ spikes in } [0,t] \mid \theta) = \dfrac{[\nu_i(\theta) t]^{k_i}}{k_i!} e^{-t\nu_i(\theta)}$

$\nu_i(\theta) = \phi(|\theta - \theta_i|_{\text{mod } 2\pi})$ tuning curve

detailed study:

Seung & Sompolinsky PNAS 90 (1993) 10749-10753

coding efficiency $\leftrightarrow$ "Fisher information"

ϕ ? $\leftrightarrow$ optimal performance $\leftrightarrow$? $\max \int d\theta\, \rho(\theta) \mathcal{F}(\theta)$

N Brunel & JPN: $I[\vec{k}, \theta] \xrightarrow{\text{large } p}$ cst. $+ \frac{1}{2} \int d\theta\, \rho(\theta) \ln \mathcal{F}(\theta)$

[saddle point method]

$$\mathcal{F}(\theta) = t \sum_i \frac{[\nu_i'(\theta)]^2}{\nu_i(\theta)} = t\, p \int d\theta_0\, r(\theta_0)\, G(\theta - \theta_0)$$

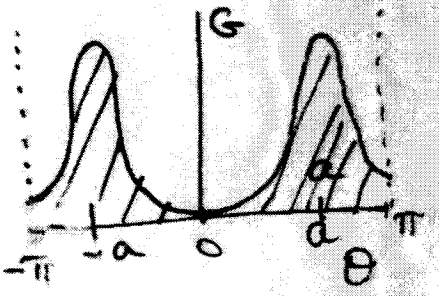
$$G(\theta - \theta_0) \equiv \frac{\phi'^2(\theta - \theta_0)}{\phi(\theta - \theta_0)}$$

Optimal choice of the θ_i's ? (of $r(\theta_0)$):

$r(.)$ such that $\int d\theta_0\, r(\theta_0)\, G(\theta - \theta_0) \propto \rho(\theta)$

For r_{opt} :

$$I = \text{cst.} + \frac{1}{2} \ln \int d\theta\, G(\theta) + \frac{1}{2} \ln tp$$



'natural' images analysis ↔ sensory coding
↘ image processing

- Barlow: early coding ↔ regularities of the environment
  ⇝ substract from the stimulus everything that can be expected.
  ⇒ understanding of visual processing ↔ statistical regularities of visual stimuli (natural images)

- statistical symmetries in images:
  - translational invariance
  - scale invariance : self-similarity
    power spectrum of images: $S(f) \sim \frac{1}{f^{2-\delta}}$ (δ small)
    (D. Field 1987)

models with translational inv. & scale inv.:

Linear, Gaussian models [Atick et al 1992, 1994; van Hateren 1992; Field 87]
Atick & Li : ⇝ wavelet analysis

but: - non Gaussian statistics, multiscaling
D. Ruderman & W. Bialek 1994
A. Turiel, G. Mato, N. Parga, JPN 1998 PhysRevLett 80:1098-1101
↳ similarity with turbulent flows ("Extended self similarity")
A. Turiel + N. Parga et al: multifractal analysis (2000, 2002)

- optimal wavelet analysis: ↳ assuming scaling prop. exact
  optimal motion wavelet derived from data
- model: multiplicative (infinitely divisible) process
  ⇝ non linear processing: ratios of wavelet coeff. associated to different scales are statistically independent.

- Most Singular Manifold (~ edges)
  image = MSM ⊕ "sources"
  (singular) (smooth)
  reconstruction of the image from the MSM alone
  Applications { images, turbulence; geophysical data

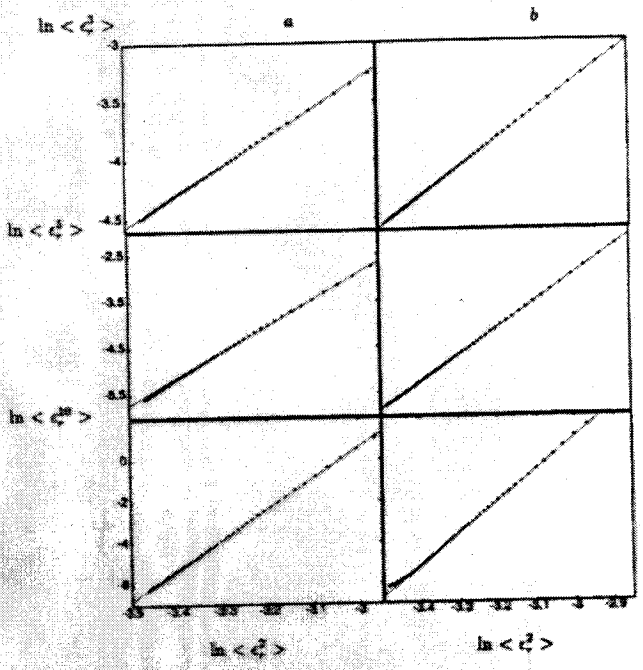


FIG. 2. Test of ESS. We plot $\ln\langle\epsilon_{l,r}^p\rangle$ vs $\ln\langle\epsilon_{l,r}^2\rangle$ for $p = 3$, 5, and 10. Data correspond to scales from $r = 8$ to $r = 64$ pixels. The effect of finite size effects can again be observed for r close to 64 pixels. (a) Horizontal direction, $l = h$; (b) vertical direction, $l = v$. The solid lines are the slope given by the calculated exponents $\rho(p, 2)$.

The difference between Eqs. (5) and (6) can also be phrased in terms of multiplicative processes [28,29]. Instead of $f_r \sim f_L$, we now have $f_r \sim \alpha f_L$, where the factor α itself becomes a stochastic variable determined by the kernel $G_{rL}(\ln\alpha)$. Since the scale L is arbitrary (scale r can be reached from any other scale r'), the kernel must obey a composition law. This stochastic variable at scale r can then be obtained through a cascade of infinitesimal processes $G_\delta \equiv G_{r,r+\delta r}$.

Specific choices of G_δ define different models of ESS. The She-Leveque (SL) [6] model corresponds to a simple process such that α is 1 with some probability $1 - s$ and is a constant β with probability s. One can see that $s = \frac{1}{1-\beta^2}\ln(\frac{\langle f_r^2\rangle}{\langle f_L^2\rangle})$ and that this stochastic process yields a log-Poisson distribution for α [30]. It also gives ESS with exponents $\rho(p, q)$ that can be expressed in terms of a single parameter (β) as follows [6]:

$$\rho(p, q) = \frac{1 - \beta^p - (1 - \beta)p}{1 - \beta^q - (1 - \beta)q}. \tag{7}$$

We have tested the model with the ESS exponents obtained with the image data set. The resulting fit for the SL model is shown in Fig. 3. Both the vertical and horizontal ESS exponents can be fitted with $\beta = 0.50 \pm 0.03$. More complex processes other than log-Poisson

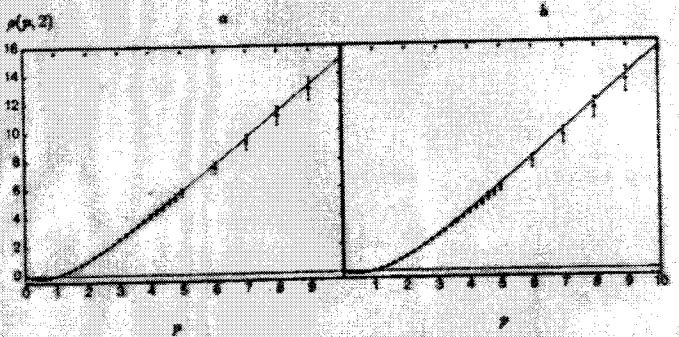


FIG. 3. ESS exponents $\rho(p, 2)$, for the vertical and horizontal variables. Each value of $\rho_l(p, 2)$ was obtained by a linear regression of $\ln\langle\epsilon_{l,r}^p\rangle$ vs $\ln\langle\epsilon_{l,r}^2\rangle$ for distances r between 8 and 64 ($l = v, h$). (a) Horizontal direction, $\rho_h(p, 2)$; (b) vertical direction, $\rho_v(p, 2)$. The solid line represents the fit with the SL model. The best fit is obtained with $\beta_v \sim \beta_h \sim 0.50$. The error bars b_p have been estimated by dividing the 45 images into 9 groups, evaluating $\rho_l(p, 2)$ for each of them, and computing the dispersion of these values. The errors grow as p increases. This is because moments of higher order are sensitive to the tail of the distribution of the local edge variance. The fit is such that the following average quadratic error, $E = \sum_p \frac{[\rho(p,2)_{exp} - \rho(p,2)_{th}]^2}{b_p}$, is minimized. We have checked that a Gaussian data set of images does exhibit ESS although it cannot be explained by the SL model.

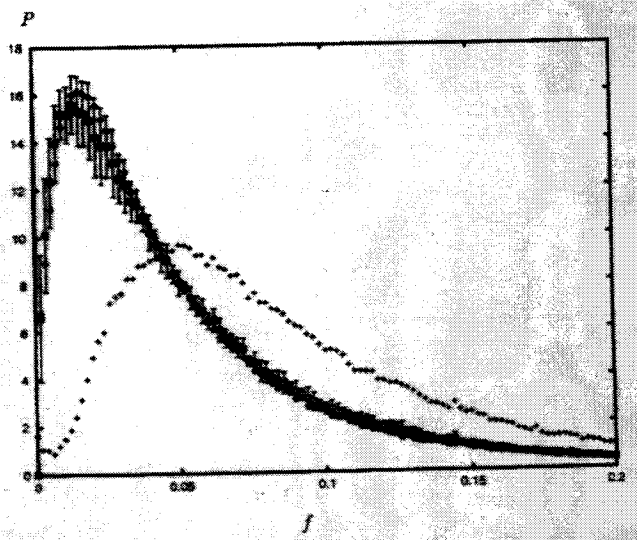


FIG. 4. Verification of the validity of the integral representation of ESS, Eq. (6) with a log-Poisson kernel, for horizontal local edge variance. The largest scale is $L = 64$. Starting from the histogram $P_L(f)$ (crosses), and using a log-Poisson distribution with parameter $\beta = 0.50$ for the kernel G_{rL}, Eq. (6) gives a prediction for the distribution at the scale $r = 16$ (squares). This has to be compared with the direct evaluation of $P_r(f)$ (diamonds). Similar results hold for other pairs of scales. The error bars have been estimated as follows: The data set was divided in nine groups, as explained in the previous figure, and the histograms at the scales L and r were computed for each group. Then for each group the histogram at scale L was used to obtain a prediction for the histogram at scale r. The differences between the predicted and computed values were squared and averaged over the groups. Its square root gives a measure of the error committed *in the prediction*, represented by the error bars. The test for the vertical case is as good as for the horizontal variable.

Antonio TURIEL

http://www.lps.ens.fr/~risc/rescomp/Antonio/

face the problem of how to obtain the optimal wavelets ϕ_r from the optimal weighted average Ψ. A possible way to do it is described in the next subsection.

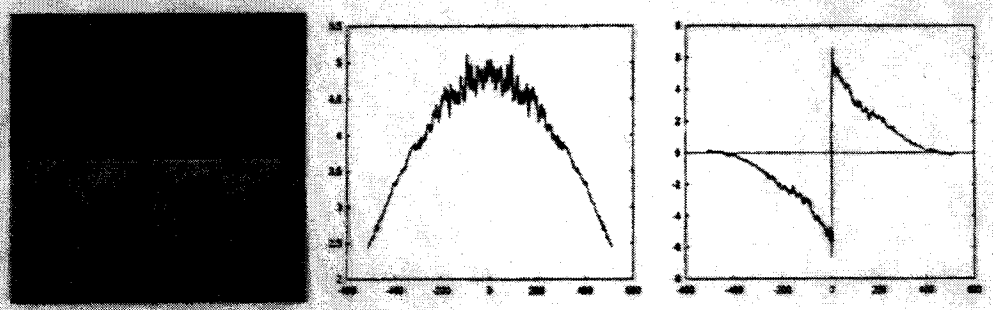


Fig. 2. Left: Gray level representation of the optimal wavelet Ψ. Middle: Horizontal cut. Right: Vertical cut

4.4 *Multiple orientations*

In order to obtain the wavelet family $\{\phi_r\}$ it is necessary to make further assumptions. The simplest guess is that they are rotated versions of the same detector, and that they are all mutually orthogonal. In [51] the theoretical derivation to extract ϕ_r from Ψ is presented. There exist in general several possible choices for the first feature detector ϕ_0 (from which all the others are obtained by simple rotation); the simplest is the one which resembles the most to Ψ. As an experimental observation $\phi_0 = \Psi$ with a great accuracy, up to $n = 8$ different orientations. So, for this image ensemble, it is possible to take the average detector Ψ as the general feature detector; we will make use of this in the following.

Before discussing the evidence in favor of the assumptions made to obtain the optimal detector (see section (6) below) we now present the main implications of the results.

5 Implications of the efficient representation

Scale invariance leads to a non-linear code: Scale invariance gives a non-linear efficient representation. Non-linearities of simple cells are indeed well-known (see e.g. [24; 8]) The efficient code has a first, linear stage where the orientation is detected. As we have seen, this linear response *is not* an efficient code. It corresponds to the wavelet coefficients and these are very correlated through

Estimation Theory

CRAMER-RAO bound and FISHER information

$$\theta \xrightarrow{Q(\underline{X}|\theta)} \underline{X} \longrightarrow \hat{\theta}(\underline{X})$$

parameter — data — estimator

- Case of unbiased estimator: $\langle \hat{\theta} \rangle = \theta$

quadratic error: $\sigma_\theta^2 \equiv \int d^p X \, Q(\underline{X}|\theta) \left(\hat{\theta}(\underline{X}) - \theta \right)^2$

CRAMER-RAO (1945): $\sigma_\theta^2 \geq \frac{1}{F(\theta)}$

$F(\theta)$ = Fisher information $= \left\langle \left(\frac{\partial \ln Q}{\partial \theta} \right)^2 \right\rangle$

Optimal bound: equality for specific cases ("efficient estimator")

Similar to uncertainty principle in Quantum Mechanics

$$\langle (\hat{\theta} - \theta)^2 \rangle \, \langle \left(\frac{\partial \ln Q}{\partial \theta} \right)^2 \rangle \geq 1$$

simple proof via Schwart ineq. $(\langle x^2 \rangle \langle y^2 \rangle \geq \langle xy \rangle^2)$

- generalizations:

* with bias $\langle \hat{\theta} \rangle \neq \theta$ $\quad \sigma_\theta^2 \geq \dfrac{\left[\frac{d}{d\theta} \langle \hat{\theta} \rangle \right]^2}{F(\theta)}$

[psychophysics: discriminability d'

measure of performance in a discrimination task θ ? $\theta + \delta\theta$

$d' \leq \delta\theta \sqrt{F(\theta)}$

* $\underline{\theta} \in \mathbb{R}^N$ covariance matrix $\Sigma(\theta)$

$$\Sigma_{jj'} = \langle (\hat{\theta}_j - \theta_j)(\hat{\theta}_{j'} - \theta_{j'}) \rangle$$

Fisher information matrix: $F_{jj'}(\theta) = \left\langle \frac{\partial \ln Q}{\partial \theta_j} \frac{\partial \ln Q}{\partial \theta_{j'}} \right\rangle$

$\Sigma \geq F^{-1}$

$(\forall u \quad u^T (\Sigma - F^{-1}) u \geq 0)$

Fisher information and Shannon information

$$F(\theta) = \left\langle \left(\frac{\partial \ln Q}{\partial \theta} \right)^2 \right\rangle = \left\langle - \frac{\partial^2 \ln Q}{\partial \theta^2} \right\rangle$$

$$= - \int d^p X \; Q(X|\theta) \; \frac{\partial^2}{\partial \theta^2} \ln Q(X|\theta)$$

Fisher information is NOT a Shannon information

However :

$\underline{X} \in \mathbb{R}^p \qquad p \to \infty$

$$I[\underline{X}, \theta] = - \int d\theta \, \rho(\theta) \ln \rho(\theta) + \frac{1}{2} \int d\theta \, \rho(\theta) \ln \frac{F(\theta)}{2\pi e}$$

Clarke & Barron 1990 — IEEE Inf. Th. 36 : 453-471

Rissanen 1996 — IEEE Inf Th 42 : 40-47

N. Brunel & JPN 1998 — Neural Computation 10 : 1731-1757
(simple derivation with saddle point techn.)

link between optimal information transfer
and optimal reconstruction

Relation easy to understand:

large number of observations

↳ $P(\theta | \underline{X})$ ~ Gaussian,

centered on Maximum Likelihood estim.
(optimal unbiased estimator for $p \to \infty$),
with variance = Fisher information.

mutual information ↔ entropy of this Gaussian pdf.

- Case $\theta \in \mathbb{R}^N$

$$p \to \infty \qquad I = - \int d^N\theta \, \rho(\theta) \ln \rho(\theta) + \frac{1}{2} \int d^N\theta \, \rho(\theta) \ln \frac{|\det F|}{(2\pi e)^N}$$

Information-Theoretic Asymptotics of Bayes Methods

BERTRAND S. CLARKE AND ANDREW R. BARRON, MEMBER, IEEE

...tract—In the absence of knowledge of the true density function, ...esian models take the joint density function for a sequence of n ...dom variables to be an average of densities with respect to a prior. ... examine the relative entropy distance D_n between the true density ... the Bayesian density and show that the asymptotic distance is ...2X log n) + c, where d is the dimension of the parameter vector. ...erefore, the relative entropy rate D_n/n converges to zero at rate ... n)/n. The constant c, which we explicitly identify, depends only on ... prior density function and the Fisher information matrix evaluated ...he true parameter value. Consequences are given for density estima... universal data compression, composite hypothesis testing, and ...-market portfolio selection.

we identify. We note that if the mixture excludes a neighborhood of the true density, then the behavior of the relative entropy is, asymptotically, of the order of the sample size; in addition, if the prior is discrete and assigns positive mass at θ_u, the relative entropy then asymptotically tends to a constant.

The relative-entropy rate between the true distribution and the mixture of distributions has been examined by Barron [4]. It is shown that if the prior assigns positive mass to the relative entropy neighborhoods $\{\theta: D(P_{\theta_u} \| P_\theta) < \epsilon\}$, $\epsilon > 0$, then

0 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 42, NO. 1, JANUARY 1996

Fisher Information and Stochastic Complexity

Jorma J. Rissanen, *Senior Member, IEEE*

Abstract—By taking into account the Fisher information and ...emoving an inherent redundancy in earlier two-part codes, ... sharper code length as the stochastic complexity and the ...ssociated universal process are derived for a class of parametric ...rocesses. The main condition required is that the maximum-...kelihood estimates satisfy the Central Limit Theorem. The ...ame code length is also obtained from the so-called maximum-...kelihood code.

Index Terms— Universal coding, universal modeling, MDL ...rinciple.

I. Introduction

THE idea of universal coding, suggested by Kolmogorov, ... is to construct a code for data sequences such that ...symptotically, as the length of the sequence increases, the ...ean per symbol code length would approach the entropy ...f whatever process in a family has generated the data. ...n the seminal works by Davisson [5], [6], as well as by ...Krichevsky and Trofimov [9], Rissanen [13], Shtarkov [19], ...nd others, this has been shown to be possible for many ...f the usual types of finite alphabet processes. Different ...niversal codes can be compared in terms of the mean code ...edundancy of the worst case process, i.e., the mean code ...ength excess over the entropy, maximized over the processes ...n the considered family. An interesting result due to Gallager ...n 1974 (unpublished lecture notes) states that the worst case ...ean redundancy for the best code equals the channel capacity, ...hen the family of processes $\{f(x^n \mid \theta)\}$, together with a prior ...(θ), is viewed as an information channel. The capacity, then, ... defined as the maximized mutual information $I_w(\Theta; X) =$

We indicate logarithm to the base two by "log" and the natural logarithm by "ln." Originally this prior was constructed by invariance arguments for the purpose of capturing the elusive idea of no prior knowledge.

Recently, these studies were carried further by Clarke and Barron [3], [4], who were able to provide a very accurate asymptotic formula for the redundancy of the code, defined by the mixture density

$$f_w(x^n) = \int f(x^n \mid \theta)\, dw(\theta)$$

namely

$$E_\theta \ln \frac{f(x^n \mid \theta)}{f_w(x^n)} = \frac{k}{2} \ln \frac{n}{2\pi e} + \ln \frac{|I(\theta)|^{1/2}}{w(\theta)} + o(1) \quad (2)$$

when the modeled processes are independent and identically distributed (i.i.d.), satisfying suitable smoothness conditions. From this, an asymptotic formula for the channel capacity follows if we take the prior as Jeffreys' prior (1).

Gradually over the years, universal coding has evolved into something that could be called *universal modeling*. The purpose is no longer restricted to just encoding of data but rather to finding optimal models, above all an optimal universal model, to be used whenever models of the data-generating machinery are needed. Universal modeling with the associated Minimum Description Length (MDL) principle for statistical inference, then, generalizes the older idea of parameter estimator in statistics [12], [15], and it incorporates the model complexity which affects all aspects of model

signal to noise ratio of the neuron. The total information is given by the sum of $J[r_i](\theta)$ for all neurons, which for large N is

$$J[\mathbf{r}] = N \int_0^{2\pi} \frac{d\phi}{2\pi} \frac{f'(\phi)^2}{f(\phi)}. \qquad [3]$$

Note that because of the isotropic distribution of the preferred directions θ_i, $J[\mathbf{r}]$ is the same for any stimulus θ in the continuum limit.

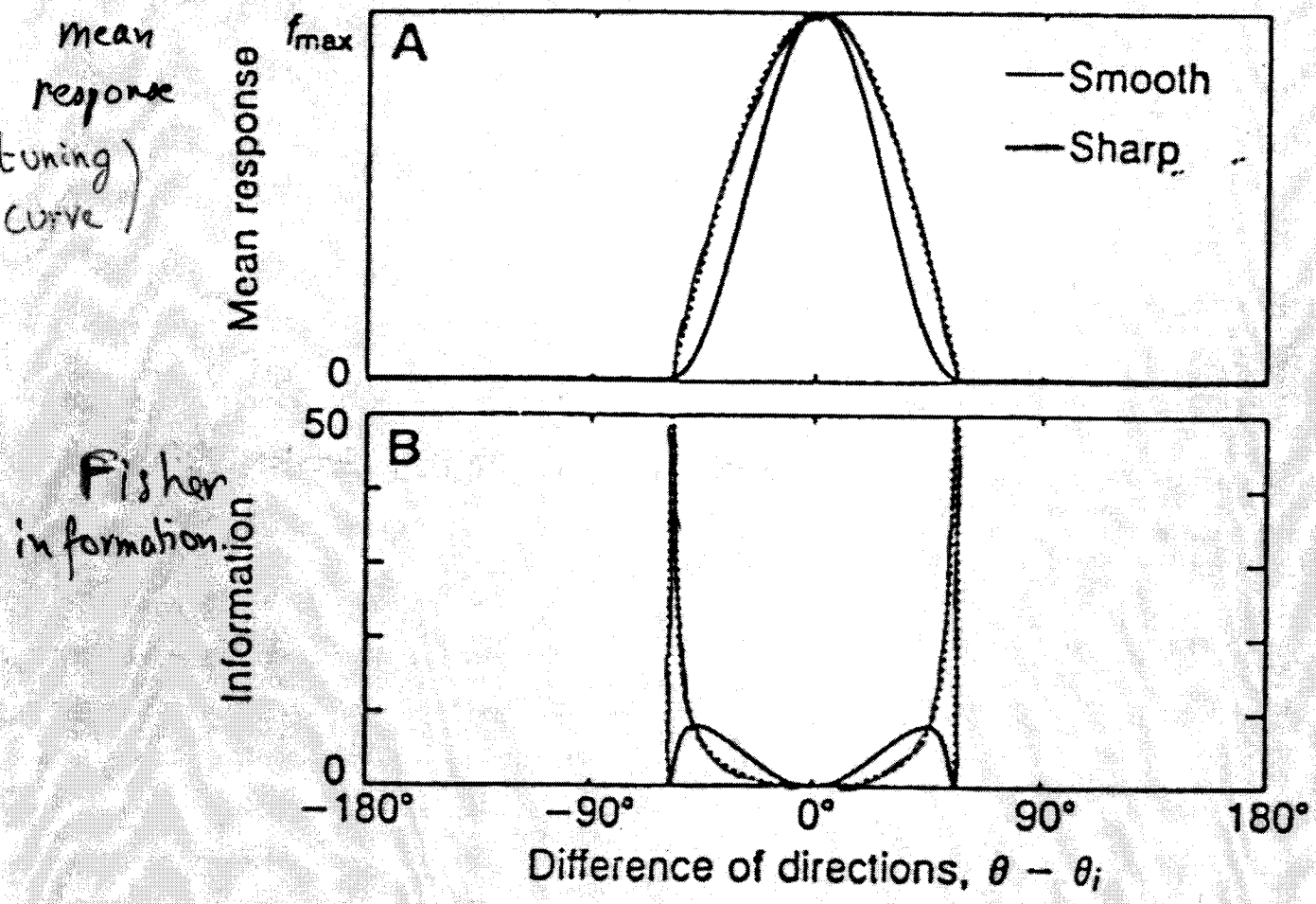


FIG. 1. (*A*) Two tuning curves of the form given in Eq. 1. Both smooth ($m = 2$, solid line) and sharp ($m = 1$, dotted line) thresholds are shown. The ratio of background to peak response is $\rho = f_{min}/f_{max} = 0.01$, and the width is $a = 1$. (*B*) Information $J[r_i](\theta)$ in neuron i as a function of $\theta_i - \theta$ for tuning curves with $a = 1$ and $\rho = 0.01$. There is no information in the neurons with $\theta_i \approx \theta$, at their maximal firing rates. For the sharp threshold population, the peak of J is at $|\theta - \theta_i| = a$ and extends well beyond the top of the figure.

Seung & Sompolinsky
1993

example: population coding

study of mutual information: N Brunel & JPN 1998

Fisher information: Seung & Sompolinsky 1993

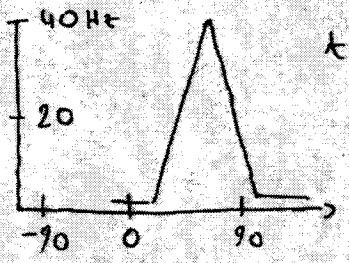
model: Poisson process

X_i = number of spike of cell i between 0 and t

θ = angle

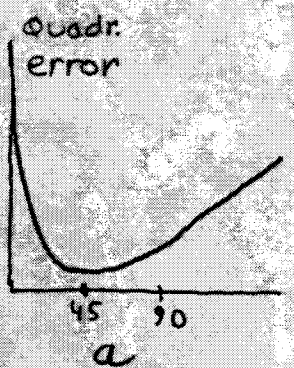
$$Q(X|\theta) = \prod_i Q_i(X_i|\theta) \qquad Q_i(X_i = n|\theta) = \frac{[t\nu_i(\theta)]^n}{n!} e^{-t\nu_i(\theta)}$$

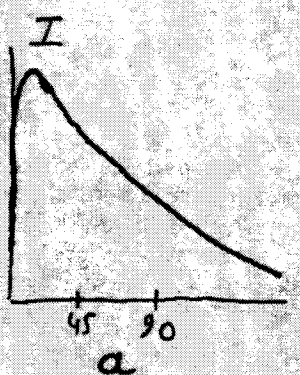
$$\nu_i(\theta) = \nu\left(\frac{\theta - \theta_i}{a_i}\right)$$



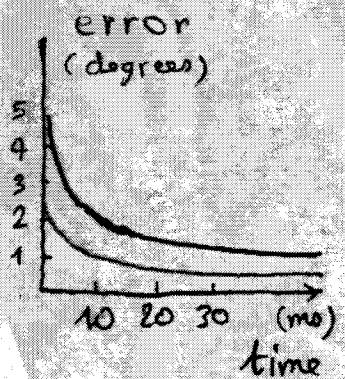
tuning curve

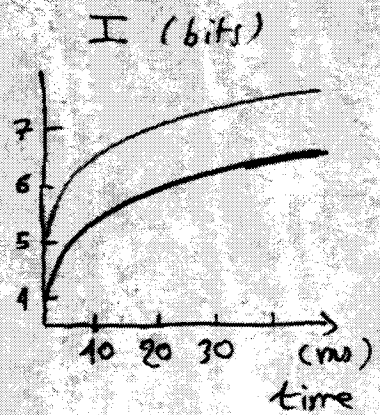
- short times: after a single spike





- p large





p = 1000 —

p = 5000 —

tuning curves

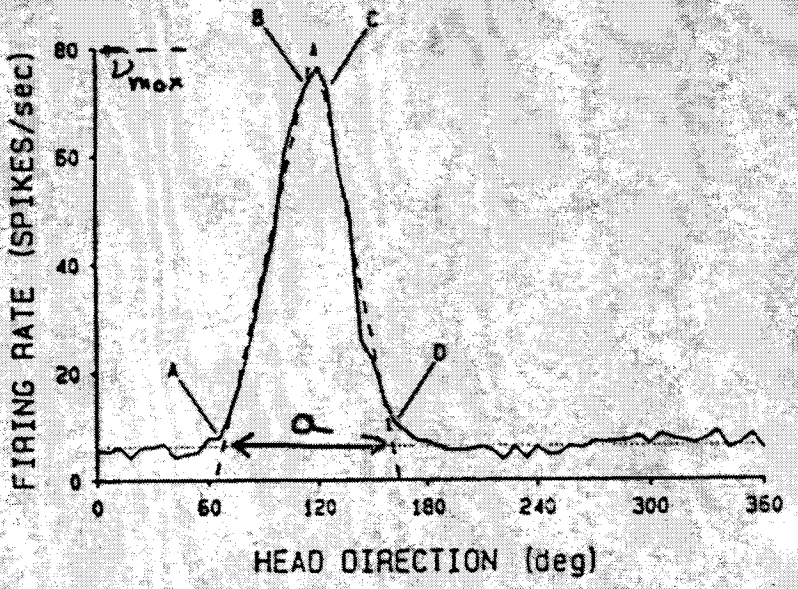


Figure 4. Summary of the triangular model of head-direction cell firing. The *solid curve* shows firing rate as a function of head direction for a typical head-direction cell. The 2 *dashed lines*, along with the base, form the triangle that was considered to best fit the raw data. The angle below the apex of the triangle is taken to be the preferred direction (118.5°). The peak firing rate is taken as the height of the triangle (80.2 spikes/sec). The directional firing range extends from the *X* intercept of the left leg of the triangle (63.5°) to the *X* intercept of the right leg of the triangle (164.6°) and is 101.1°. The left leg of the triangle was obtained from a least-squares fit of the 9 data points between and including points *A* and *B*. The right leg of the triangle was obtained from a least-squares fit of the 7 data points between and including points *C* and *D*. The magnitude of the background firing rate (6.26 AP/sec) is indicated by the horizontal *dotted line*. The background firing rate was taken as the mean firing rate for all data points more than 18° counterclockwise from the *X* intercept of the left leg of the triangle and more than 18° clockwise from the *X* intercept of the right leg of the triangle (in this case, 39 points). The asymmetry of the triangle was defined as the absolute value of the left leg slope divided by the right leg slope.

Taube et al 1990 J. of Neuroscience 10: 42

<Error>

I [first spike ; stimulus]

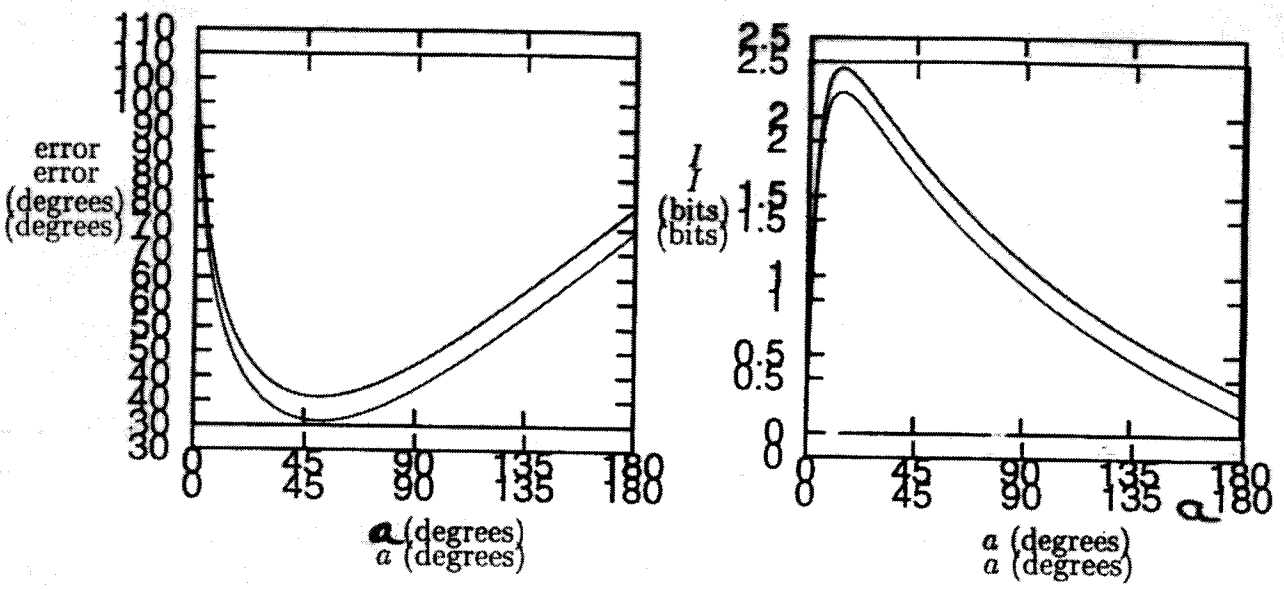
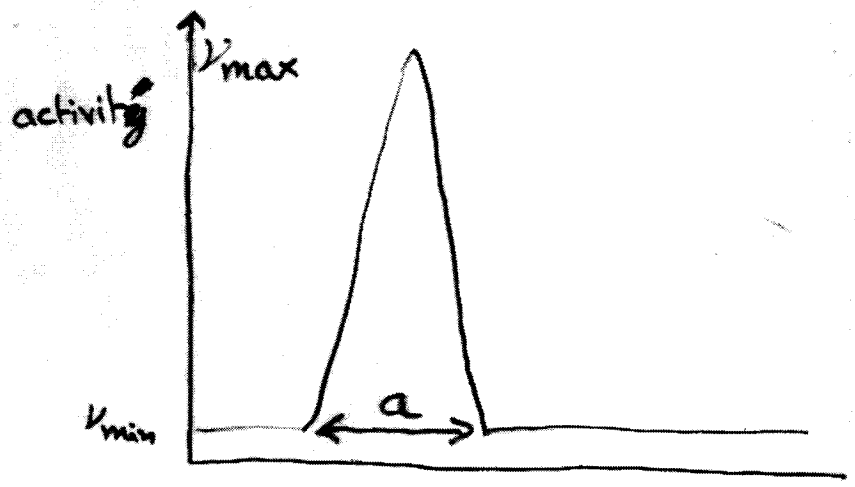


Figure 2: Left: SD of the reconstruction error after a single spike, as a function of a. Right: mutual information between the spike and the stimulus as a function of a. Note that minimizing the SD of the reconstruction error is in this case different than maximizing the mutual information.

~~The mutual information, on the other hand, is~~
The mutual information, on the other hand, is

[N. Brunel & JPN 1998]



N. Brunel & JPN 1998

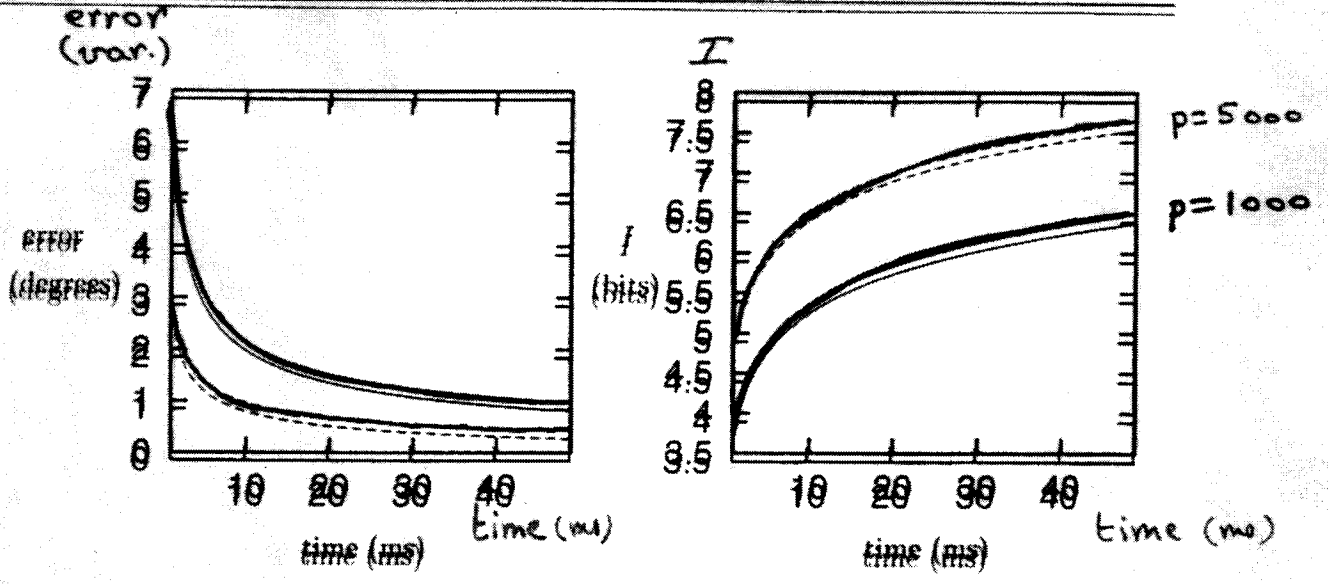


Figure 3: (Left) Minimal reconstruction error as given by the Cramer-Rao bound for $N = 1000$ (full curve), $N = 5000$ (dashed curve) postsubiculum neurons, using data from Taube et al. (1990) as a function of time. (Right) Mutual information for $N = 1000$ (full curve), $N = 5000$ (dashed curve), using the same data and equation 3.10.

Under global constraints, one may expect each neuron to contribute in the same way to the information, that is, $(\nu_{max}/a)(1 - 1/\alpha) \ln a$ is constant. This would imply that the width a increases with ν_{max}. Figure 9 of Taube et al. (1990) shows that there is indeed a trend for higher firing rate cells to have wider directional firing ranges.

We can now insert the distributions of parameters measured in Taube et al. (1990) in equation 5.8 to estimate the minimal reconstruction error that can be done on the head direction using the output of N postsubiculum neurons during an interval of duration t. It is shown in the left part of Figure 3. Since we assume that the number of neurons is large, the mutual information conveyed by this population can be estimated using equation 3.9. It is shown in the right part of the same figure. In the case of $N = 5000$ neurons, the error is as small as one degree even at $t = 10$ ms, an interval during which only a small proportion of selective neurons has emitted a spike. Note that one degree is the order of magnitude of the error made typically in perceptual discrimination tasks (see, e.g., Pouget & Thorpe 1991). During the same interval, the activity of the population of neurons carries about 6.5 bits about the stimulus. Doubling the number of neurons or the duration of the interval divides the minimal reconstruction error by $\sqrt{2}$ and increases the mutual information by 0.5

Schoups et al
Nature 412 2 Aug. 2001

in this type of learning to an improved neuronal performance of trained compared to naive neurons. Improved long-term neuronal performance resulted from changes in the characteristics of orientation tuning of individual neurons. More particularly, the slope of the orientation tuning curve that was measured at the trained orientation increased only for the subgroup of trained neurons most likely to code the orientation identified by the monkey. No modifications of the tuning curve were observed for orientations for which the monkey had not been trained. Thus training induces a specific and efficient increase in neuronal sensitivity in V1.

The psychophysics of early perceptual learning has been well studied; however, researchers are only beginning to understand the neurophysiological correlates of perceptual learning. Primary somatosensory, motor and auditory cortex show learning-dependent changes in their topographical organization[9–11]. In the visual system, however, early plasticity is demonstrated only by lesion-dependent reorganizations[12,13]. Improvement in motion discrimination, which transfers to other retinal stimulus positions, is accompanied by short-term improvements in neuronal sensitivity in the higher-order middle temporal and medial superior temporal areas within a single session, but does not extend across sessions[14,15].

We trained two monkeys to identify the orientation of a small grating (Fig. 1). The performance of monkeys, like that of human subjects[7], improved markedly with training, reaching a threshold as low as 0.6–1.2° after several months. The improvement was specific for both stimulus position and orientation. This specificity provided us with an internal control: instead of comparing data between monkeys, we could compare different populations of

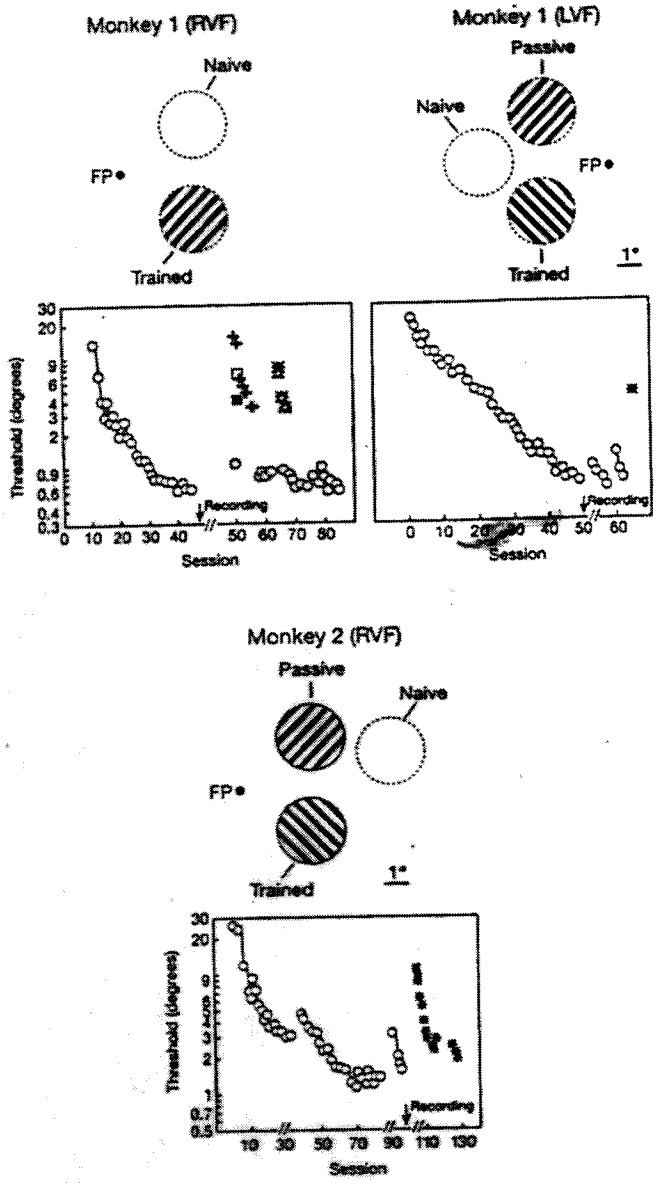


Figure 1 Behavioural performance and specificity for position and orientation. Learning curves for orientation identification for monkey 1 (two hemispheres) and monkey 2 (RVF). Above each learning curve, a schematic representation of the stimulus positions and reference orientation is given for each of the hemispheres studied. Thresholds for the trained orientation and position decreased on consecutive daily sessions (open blue circles). After recording, transfer was tested to the other oblique orientation (filled blue squares) at the same stimulus position, and to other stimulus positions (green, naive position; red, passively stimulated position). Thresholds at these other stimulus positions were similar for the two oblique orientations (green stars and triangles; red plus signs and open square), and were 7–12 times larger than for the trained position and orientation. FP, fixation point.

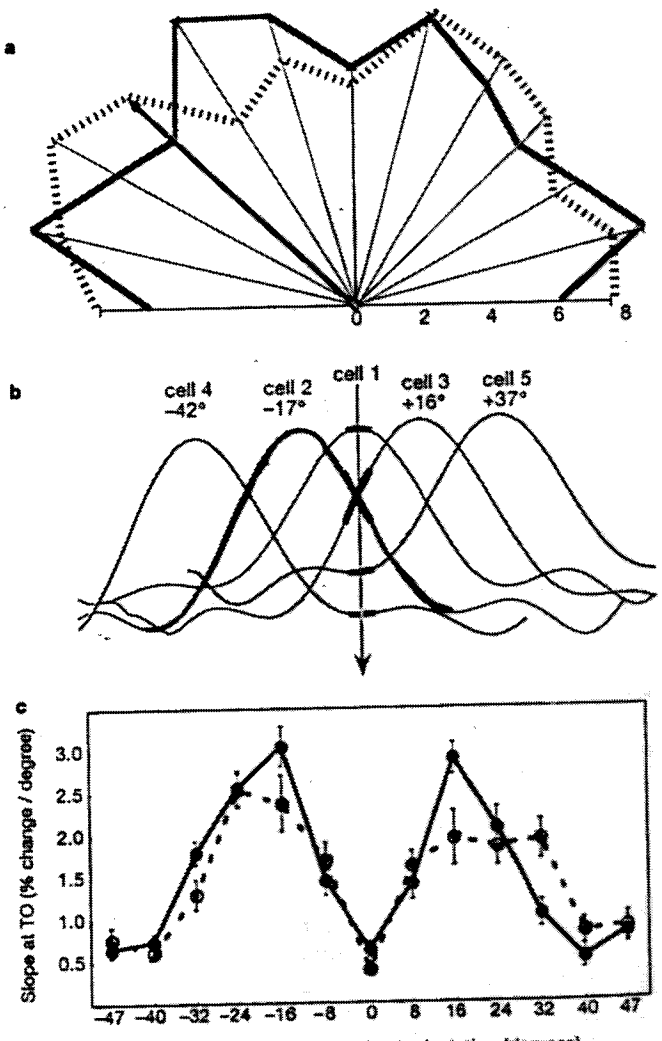


Figure 2 Neuronal responses. **a**, Distribution (%) of preferred orientations in trained (solid red line) and naive neurons (dashed blue line). Data are from all cell layers. The green arrow indicates the trained orientation. **b**, Orientation tuning curves of five sample neurons. **c**, Slope measured at trained orientation (TO) for trained (solid red line) and naive neurons (dashed blue line). Neurons are classified according to the angle between the preferred and trained orientation. Neurons tuned to the trained orientation are shown in the centre, groups with preferred orientation clockwise from trained orientation are to the left; those anticlockwise to the right. Data (± s.e.m.) are from layers 2–3 and 5–6 combined.

smooth case

ex.: population coding with smooth enough tuning curve

$P \to \infty$

$$I(\theta, V) \simeq H[\theta] + \int d^N\theta\, \rho(\theta)\, \frac{1}{2} \ln \frac{|\det F(\theta)|}{2\pi e}$$

$F(\theta)$ = Fisher information

$$F_{ij}(\theta) = \left\langle -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln P(X|\theta) \right\rangle_\theta$$

$$\longrightarrow \quad I \simeq \frac{N}{2} \ln P$$

Clarke & Barron 1990; Rissanen 1996; Brunel and Nadal 1998

* Population coding

Poisson processes with smooth tuning curves

[Seung & Sompolinsky 1993 : Fisher information]

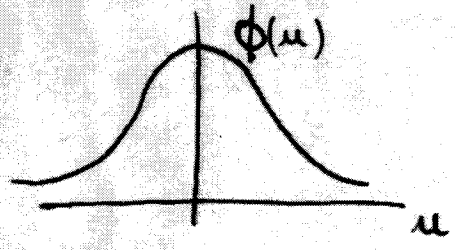
[Brunel & Nadal 1998 : mutual information]

$N=1$ $\quad P_i(X_i|\theta) = \dfrac{[\nu_i(\theta)\, t]^{V_i}}{V_i!} \exp - t\nu_i(\theta)$

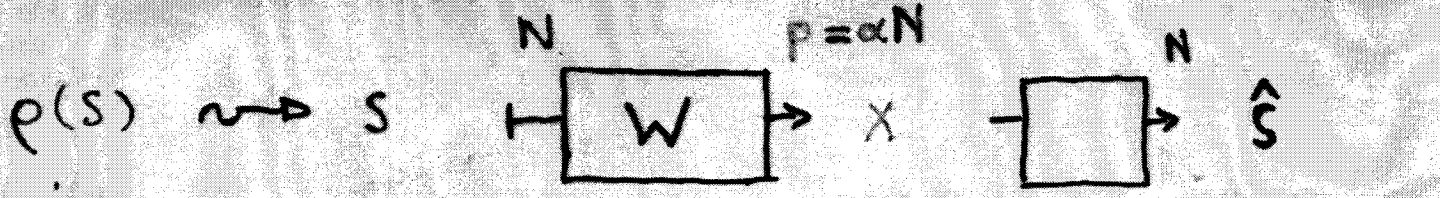
tuning curve $\nu_i(\theta) = \phi\left(\dfrac{\theta - \theta_i}{a_i}\right)$

$$F(\theta) = t \sum_{i=1}^{P} \frac{[\nu_i'(\theta)]^2}{\nu_i(\theta)} \simeq t\rho \int \frac{d\theta'}{a^2} r(\theta') \frac{[\phi'(\frac{\theta-\theta'}{a})]^2}{\phi(\frac{\theta-\theta'}{a})}$$

(case $a_i = a$)



Bayesian approach to parameter estimation



prior — parameter — model — data — estimator generalization

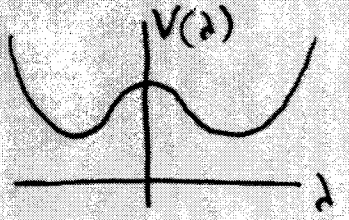
exples: unsupervised learning

$S = \underline{B} \in \mathbb{R}^N \qquad X = \{\underline{I}^1, \dots, \underline{I}^P\} \qquad \underline{I}^\mu \in \mathbb{R}^N$

$\rho(\underline{B}) : \underline{B}^2 = 1$

$$P[X \mid \underline{B}] = \prod_{\mu=1}^{P} \frac{1}{\sqrt{2\pi}^N} \exp - \frac{1}{2} \underline{I}^{\mu 2} - V(\lambda^\mu)$$

$$\lambda^\mu = \underline{I}^\mu \cdot \underline{B}$$



Biehl Mietzner 94; Watkin N. 94; Reimann Van den Broeck 96; Buhot Gordon 98; Herrschkowitz N. 98

supervised learning

$$\text{Data} = \{(\underline{I}^\mu, \sigma^\mu), \mu = 1, \dots, P\}$$
$$\sigma^\mu = \text{sgn}(\lambda^\mu)$$

- $N \to \infty$ $\quad \alpha = \frac{P}{N}$ fixed $\quad$ replica calculations
- N fixed $\quad P >> N$ $\quad$ Amari Murata 92; Haussler Opper 97; Herrschkowitz JPN 98.

mutual information $I[X, S]$ =

- information conveyed by the data over the parameter
- cumulative relative entropy loss (Bayes risk)

(Haussler & Opper 1997 The Annals of Statistics 25 p 2451-2492)

$\theta \longmapsto X = \{X_1, \dots, X_p\}$

- linear upper bound $I \leq \gamma p$ [Herschkowitz & Nadal 1998]
  - * if $d_{VC} = \infty$ $I \propto p$
  - * if $d_{VC} < \infty$:
- θ discrete (e.g. $\theta_j = \pm 1$)

  $I \leq$ entropy$[\theta] = -\sum_\theta p(\theta) \ln p(\theta) \equiv H[\theta]$

  $p \to \infty$ $I \to H[\theta]$ (exponentially)

  [Opper, Haussler]
- $\theta \in \mathbb{R}^N$

  <u>smooth</u> case $p \to \infty$ $I \sim \frac{N}{2} \ln$[fisher information] $\sim \frac{N}{2} \ln p$

  [Clarke & Barron 1990, Rissanen 1996, Brunel & Nadal 1998]

  ex.: population coding with smooth tuning curves

  <u>non smooth</u> case $p \to \infty$ $I \sim r N \ln \frac{P}{N}$ $r > \frac{1}{2}$

  ex.: binary perceptron: $r = 1$ [Nadal & Parga 1994]

  binary outputs: $I \leq$ growth function $\simeq d_{VC} \ln p$

  [Nadal & Parga 1994; Opper 1995]

  ($\Rightarrow d_{VC} \geq r N$)

  general case: Haussler & Opper 1997

  Herschkowitz & Nadal 1998; Herschkowitz & Opper 1999

  <u>Rem.</u>: strong correlations between the V_i's may lead to I finite.
- $N = \infty$ (θ = function) [Opper 1998]

  $p \to \infty$ $I \simeq N_{eff}(p) \ln p$

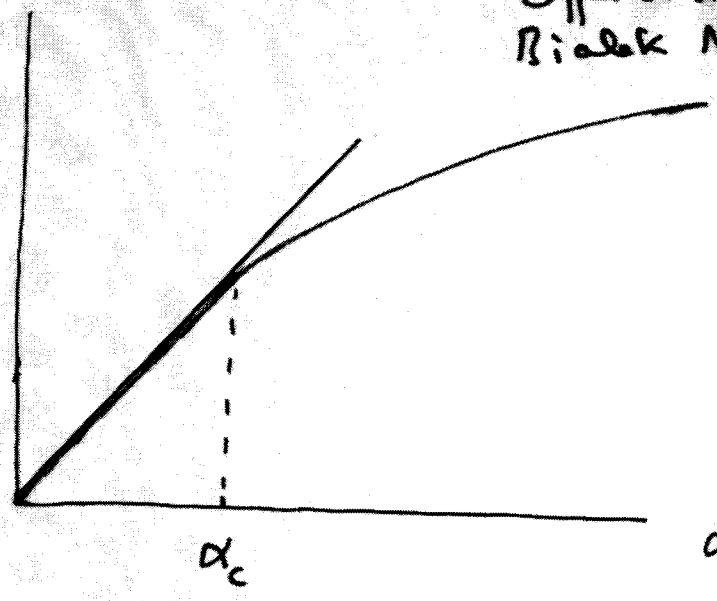
  smooth case: $N_{eff} \sim \ln p$ $I \sim (\ln p)^2$

  non smooth $N_{eff} \sim p^\mu$, e.g. $I \sim \sqrt{p}$

non parametric case ($N = \infty$)

$I[X;\theta]$ $\qquad d_{VC} \sim \sqrt{p}$ $\qquad I \sim p^{\gamma}$ $\gamma < 1$

Oller & Haussler 1995
Rissanen Nemenman Tishby 2000



* linear regime

$I \leq p\, I_0$ (exact bounds D. Herschkowitz JPN 1999)

$I = p\, I_0$ might be achieved for $\alpha < \alpha_c$

- redundancy reduction (Barlow)
  Independent Component Analysis (JPN & Parga 1994)
- retarded learning ; memory
  (critical capacity α_c , VC dimension)

* asymptotic regime $\qquad P \gg N$

$I \sim a \ln p$ $\qquad a \leftrightarrow$ smoothness of $Q(x|\theta)$

smooth case : $a = 1/2$
perceptron : $a = 1$

generalization $\varepsilon_g \searrow$ as α^{-2a}

smooth case: $I \simeq H[\theta] + \int d\theta\, p(\theta)\, \frac{1}{2} \ln \frac{F(\theta)}{2\pi e}$ (here for $N = 1$)

(Clarke, Barron 90; Rissanen 96; Brunel & N. 98)

$F(\theta) =$ "Fisher information" $= \left\langle - \frac{\partial^2 \ln Q(x|\theta)}{\partial \theta^2} \right\rangle_\theta$

Cramer Rao: $\min[\text{Quadratic Error}] \geq \frac{1}{F(\theta)}$.

$p(.) \rightarrow \Theta = \begin{Bmatrix} \theta_1 \\ \vdots \\ \theta_N \end{Bmatrix}$ $\xleftarrow{\text{mutual information}}$ $X = \begin{Bmatrix} x^1 \\ \vdots \\ x^P \end{Bmatrix}$ Redundancy ↓

capacity: $\max_{\rho} I[\Theta; X]$

adaptation: $\min_W R$ (Barlow) (Atick & Li, ...)

$\max_W I$ infomax (Stein 67, Laughlin 81, ... van Hateren, Linsker ...)

Generic behaviour of $I[\Theta; X]$