

Neural Coding of Categories:  
Information Efficiency and Optimal Population Codes

Laurent Bonnasse-Gahot<sup>1,†</sup> and Jean-Pierre Nadal<sup>1,2</sup>

(1) Centre d'Analyse et de Mathématique Sociales  
(CAMS, UMR 8557 CNRS-EHESS)

Ecole des Hautes Etudes en Sciences Sociales  
54 bd. Raspail, F-75270 Paris Cedex 06

(2) Laboratoire de Physique Statistique  
(LPS, UMR 8550 CNRS-ENS-Paris 6-Paris 7)

Ecole Normale Supérieure  
24 rue Lhomond, F-75231 Paris Cedex 05

J Comp Neurosci. 2008 Aug;25(1);169-87

Submitted May 2007.

Revised October 2007.

Accepted December 2007.

† corresponding author - email: lbg@ehess.fr -  
tel: 01 49 54 22 07 - fax: 01 49 54 21 09

## Abstract

This paper deals with the analytical study of coding a discrete set of categories by a large assembly of neurons. We consider population coding schemes, which can also be seen as instances of *exemplar models* proposed in the literature to account for phenomena in the psychophysics of categorization. We quantify the coding efficiency by the mutual information between the set of categories and the neural code, and we characterize the properties of the most efficient codes, considering different regimes corresponding essentially to different signal-to-noise ratio. One main outcome is to find that, in a high signal-to-noise ratio limit, the Fisher information at the population level should be the greatest between categories, which is achieved by having many cells with the stimulus-discriminating parts (steepest slope) of their tuning curves placed in the transition regions between categories in stimulus space. We show that these properties are in good agreement with both psychophysical data and with the neurophysiology of the inferotemporal cortex in the monkey, a cortex area known to be specifically involved in classification tasks.

**Keywords:** categorization, population coding, exemplar models, mutual information, inferotemporal cortex.

## 1 Introduction

Categorization is one of the most fundamental abilities of cognition (Harnad, 2005) and it has always been a major source of interest and attention among philosophers and scientists. It has generated a huge amount of work in various areas such as psychology, (psycho)linguistics, philosophy, artificial intelligence, machine learning, and, more recently, neuroscience. Although studies on the neural basis of categorization are quite recent, they have known a rapid and significant development over the past decade (Ashby and Spiering, 2004), involving neurophysiology (Sigala and Logothetis, 2002; Freedman et al., 2003), neuropsychology (Humphreys and Forde, 2001) and neuroimaging (Jiang et al., 2007). Theoretically, some neural models of object categorization have been posited (Riesenhuber and Poggio, 2000), but to our knowledge no studies have analytically investigated the coding efficiency of a neural population with respect to categorization.

Actually, different communities do not give exactly the same meaning to the word ‘category’. In this paper, we typically have in mind *perceptual categories* as opposed to *conceptual categories*, such as ‘chair’ as opposed to ‘furniture’. Moreover, one may distinguish two cases: what we may call ‘macroscopic’ categories (animal, human, vegetable, face, fruit...), and ‘microscopic’ categories – the categories within a macro category: a system of vowels (for example the 16 vowels of French), or a set of familiar faces, or a set of trees... In this paper we address the issue of coding efficiency, hence of the link between neural coding and classification error, when the discrimination between two or more categories can be a difficult task, in particular when a given stimulus might belong to (at least) two different categories – which is typical of microcategories. This is why our main center of interest (but not the only one) will be the case of such overlapping categories.

A strategy widely used in the brain consists in encoding information by large assemblies of neurons, each cell being selective to a specific range of stimuli. Well-known examples of such a population coding are given by the representation of movement direction in the primate motor cortex (Georgopoulos et al., 1986), the head-direction cells in rats (Taube et al., 1990), or the encoding of faces in the inferotemporal cortex of the monkey brain (Young and Yamane, 1992). In each case, the stimulus lies in a continuous space (in the simplest case of head-direction cells, the stimulus is an angle in the interval  $[0, 2\pi]$ ) and each cell has a tuning curve centered on a preferred stimulus (an angle for which the cell responds the most), with a more or less large width. However, the information on a given stimulus is contained in the activity of the whole neural assembly – not only by the cells having this angle as preferred stimulus. One can draw several parallels between models of population coding and models in other domains, giving additional motivation for studying the population coding of categories. First, one can note the similarity between population codes and radial basis functions (Poggio, 1990; Poggio and Girosi, 1990), which have been extensively used for data classification. Second, and more interestingly, a family of models called *exemplar models* (Hintzman, 1986; Nosofsky, 1986; Kruschke, 1992) have been proposed as general models of perception and categorization. Initially built in order to account for psychological data, and inspired from the ‘connexionist’ perspective with no direct link to neurobiology, these models can actually be seen as typical instances of population coding schemes. The neural interpretation of exemplar models has already been seen (Vogels, 1999; Sigala and

Logothetis, 2002; Sigala, 2004), and we will give below an explicit formulation within the framework of population coding modeling (the main idea is that each preferred stimulus is the analog of an exemplar).

Although population coding has been deeply studied for parameter or density estimation (Seung and Sompolinsky, 1993; Brunel and Nadal, 1998; Pouget et al., 1998; Averbek et al., 2006) (the task being to infer the stimulus given the neural code), no theoretical study has addressed the question of the population coding of categories. For density estimation or stimulus discrimination, the optimal distribution of the preferred stimuli of the coding cells is expected to follow the distribution of stimuli. For a classification task, we expect a different result: intuitively, the crucial regions are the class boundaries, so that more cells should be allocated to these regions. This intuition corresponds to standard results in pattern classification and machine learning (Duda and Hart, 1973; Minsky and Papert, 1969), as exemplified by the popular *Support Vector Machine* (SVM) approach to classification (Cortes and Vapnik, 1995; Schölkopf et al., 1999). Indeed, a SVM fully characterizes the class boundary by identifying, in the available data, a subset of exemplars (the so-called support vectors) which are the closest to the class boundary.

Our study seeks notably to check and quantify this general idea of allocating more resources to the boundaries, within the context of neural coding. To do so, we will make use of information theory (Blahut, 1987; Cover and Thomas, 2006) which is a relevant framework for studying information processing in the field of computational neuroscience (see e.g Rieke et al., 1997; Dayan and Abbott, 2001, and references therein). More specifically, we will characterize the information content associated with a neural code for discrete categories, and derive quantitative and qualitative properties of the most efficient neural codes. In the case of smoothly overlapping categories, in the limit of a very large number of cells with smooth tuning curves, we show that the information efficiency essentially depends on the ratio between two Fisher information values: in the numerator, one characterizing the uncertainty in inferring the category given the stimulus, and in the denominator, the (usual) Fisher information characterizing how the neural activity is specific to a given stimulus. Then the main consequence is that, in an optimized code, the (usual) Fisher information, *at the population level*, should be the greatest between categories. This is achieved by having more cells with the stimulus-discriminating parts (steepest slope) of their neuronal tuning curves placed in the transition regions

between categories in stimulus space. We will show that this main qualitative result holds also with weaker hypothesis (non smooth tuning curves), and discuss how it is modified under lower signal-to-noise ratio (e.g. in the case of rapid processing based on the first spikes).

We will show that our results find support in available empirical data from both psychophysics – in different domains: object recognition, speech recognition – and neurophysiology – in particular those recently obtained on the role in categorization of the inferotemporal cortex in monkeys.

This paper is organized as follows. Section 2 introduces the general framework, specifies the population coding scheme, and motivates the use of information theory for our analysis. Section 3 provides the main quantitative result of our paper, giving the mutual information between the activity of a neuronal population and a set of discrete categories, in the limit of a large population size. We first investigate the case of smooth tuning curves and then show that the obtained result is more general by deriving a qualitatively similar result in a simple case of sharp tuning curves. Cases that set aside this result, namely ill-defined classes and high-noise regimes, are also studied. Finally, in section 4, we discuss the predicted psychophysical consequences, and confront our result with the neurophysiology of the inferotemporal cortex of the monkey and finally present the concluding remarks. Technical details are provided in Appendices.

## 2 Methods

### 2.1 General Framework

#### 2.1.1 Feedforward Processing

We assume given a discrete set of classes/categories,  $\mu = 1, \dots, M$  with probabilities of occurrence  $q_\mu \geq 0$ , so that  $\sum_\mu q_\mu = 1$ . Each category is characterized by a density distribution  $P(\mathbf{x}|\mu)$  over the input (sensory) space. Vowels, colors or faces constitute good examples of such categories. A sensory input  $\mathbf{x} \in \mathbb{R}^K$  elicits a response  $\mathbf{r} = \{r_1, \dots, r_N\}$  from a population of stimulus-selective neurons. An estimate  $\hat{\mu}$  of  $\mu$  (or an estimate of some function of  $\mu$ ) might then be extracted from the observation of the neural activity  $\mathbf{r}$ . This processing chain can be summarized

as follow:

$$\mu \rightarrow \mathbf{x} \rightarrow \mathbf{r} \rightarrow \hat{\mu}$$

Such multi-stage feedforward scheme, which is prototypical of exemplar models (Nosofsky, 1986; Kruschke, 1992), is the basis of models of object recognition and categorization (Riesenhuber and Poggio, 2000; Sigala, 2004). It has recently found some support in monkey neurophysiology, with the identification of an assembly of stimulus-selective neurons in the inferotemporal cortex (Op de Beeck et al., 2001; Thomas et al., 2001; Freedman et al., 2003), that feeds higher-order cortical areas (such as the prefrontal cortex: see Freedman et al., 2001, 2003) involved in more behaviorally related tasks.

### 2.1.2 The Input Space

We consider  $\mathbf{x}$  as the stimulus, that is,  $\mathbf{x} = \{x_1, \dots, x_K\}$  lives in some  $K$ -dimensional (continuous) space given by the sensory system. For example, in the case of phonemes one may think of  $\mathbf{x}$  as belonging to the 2-d space defined by the two first formants F1 and F2. A vowel, say /i/, is never produced the same way, even by a same speaker, hence ‘occupies’ a certain location in the considered space. This is modeled by a density distribution  $P(\mathbf{x} = \{F1, F2\} \mid /i/)$  that defines the category /i/. As an illustration, figure 1 shows the realizations of ten vowels of American English in the F1-F2 space. Note that the overlap between categories is not a mere artefact of the projection of the signal onto a 2d space: for instance, an utterance of /ɔ/ may indeed be perceived as /ɔ/ or /ɑ/. In the case of the perception of color, the relevant space might be the 3-d space defined by hue, saturation and value. In the case of a set of familiar faces, it is not clear what is the relevant space, but one might expect its dimension to be rather large. However, even in such case, that is more generally when one would expect many dimensions to be involved in the coding of the stimuli, typical psychophysical data show that each dimension does not play the same role. On the neurophysiological side, several experimental investigations have shown that cells in the inferotemporal cortex of the monkey brain are tuned to the dimensions relevant for categorization (Sigala and Logothetis, 2002; Kiani et al., 2007). The relevant dimensions might be imposed by the experimental protocol, as in Sigala and Logothetis (2002) involving line drawings of faces that lie in a four dimensional space (eye height, eye separation, nose length and mouth height). These findings support the idea that the relevant input space might not be what is directly given

by the sensory cells, but the result of a (possibly nonlinear) pre-processing, selecting a rather small number of relevant dimensions. This is indeed a common assumption in the context of perceptual categorization (see, e.g., Palmeri and Gauthier, 2004, for a review). The hypothesis that one of the role of encoding is to project the sensory stimulus onto a small dimensional space – e.g. through principal component analysis or other multidimensional scaling algorithms – is also common in the computational literature, and in particular in the context of exemplar-based models (Nosofsky, 1986; Kruschke, 1992).

In our general framework, we do not make any explicit hypothesis on how the space to which  $\mathbf{x}$  belongs is constructed. However, in this paper, for explicit computations we will assume that it is a not too large dimensional space consisting only of dimensions that are relevant for categorization. Technically, we will assume that the number of dimensions of the input space,  $K$ , is small compared to the size of the neuronal population,  $N$ . We will come back to this issue of dimensionality at the end of the paper.

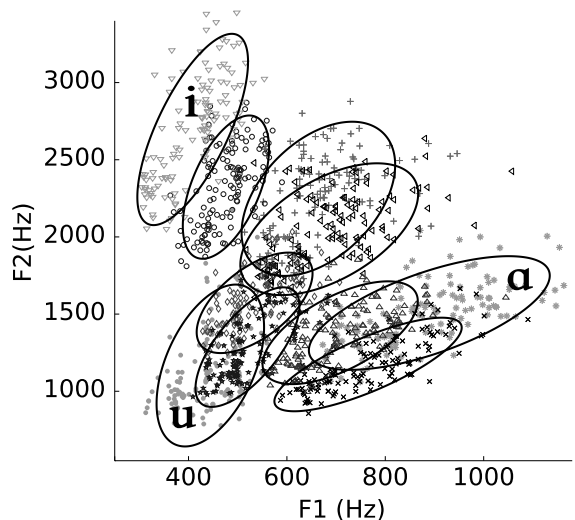


Figure 1: Representation in the F1-F2 space of ten American English vowels. Data are taken from Hillenbrand et al. (1995). To guide the eye, an ellipse is used to fit each vowel category, and for the sake of visibility, only the three corner vowels /i/, /u/ and /a/ are labeled. (data available at <http://homepages.wmich.edu/~hillenbr/voweldata.html>)

### 2.1.3 Population Coding of Categories

Although we restrict the input space to dimensions relevant for the categorization task, we do not assume  $\mathbf{x}$  to be specifically related to the categories themselves – other brain areas may use the same input for other tasks. In contrast, the main hypothesis on the neural representation  $\mathbf{r}$  is that it is specifically related to the coding of the categories. We furthermore assume that this representation is the basis for perception, hence its properties are what is revealed in psychophysical experiments. As a consequence, whenever appropriate, we will call *perceptual space* the space defined by the output of the neuronal population.

We consider an explicit family of coding schemes in the same spirit as those that have been already studied in the context of population coding, as well as in the spirit of the exemplar models mentioned above. We assume that the neuronal population has an activity  $\mathbf{r} = \{r_1, \dots, r_N\}$ , where  $r_i$  is the activity of neuron  $i$ , that depends only on the sensory input  $\mathbf{x}$  through some probability distribution  $P(\mathbf{r}|\mathbf{x})$ . Hence, given a category  $\mu$ , the neural activity is given by

$$P(\mathbf{r}|\mu) = \int d^K \mathbf{x} P(\mathbf{x}|\mu) P(\mathbf{r}|\mathbf{x}) \quad (2.1)$$

Typically, we have in mind a population code where each cell  $i$  is stimulus-selective, with a mean response characterized by a *tuning curve*  $f_i(\mathbf{x}) = R_i f(\mathbf{x}, \mathbf{x}_i, \mathbf{a}_i)$  that peaks at its *preferred stimulus*  $\mathbf{x}_i$  with a maximal rate  $R_i$ , and varies on a scale given by some parameter  $\mathbf{a}_i$  (the width of the tuning curve). For a one-dimensional sensory input, one may write:

$$f_i(x) = R_i f\left(\frac{x - x_i}{a_i}\right). \quad (2.2)$$

In the language of exemplar models, the  $\mathbf{x}_i$ 's are the stored exemplars which constitute the basis of the neural map. The width of the tuning curves are playing the same role as (the inverse of) the *attentional factors* used in exemplar-based models (Nosofsky, 1986; Kruschke, 1992) in order to weight each dimension according to its relevance – in particular unattended dimensions correspond to attributing zero weight to these dimensions. Note that, initially, a global scaling was proposed (Nosofsky, 1986; Kruschke, 1992), which would correspond to a situation where the receptive fields have, e.g., an ellipsoidal shape, but which is the same for every cell (*ie* the width  $\mathbf{a}_i$  would not depend on  $i$ ).

One can note that if the coding cells are topologically ordered according to their pre-



ferred stimulus  $\mathbf{x}_i$ , one has then a *sensory-topic* (*visuotopic*, *somatotopic*, *phonotopic*, ...) coding – hence a category with  $\mathbf{x}^\mu$  as sensory prototype leads to the activation of nearby cells with preferred stimulus  $\mathbf{x}_i$  close to  $\mathbf{x}^\mu$ .

A simple class of neural activities that we consider is the one where, given  $\mathbf{x}$ , the  $r_i$ 's are independent random variable (corresponding to a situation with no correlations between cells given the input):

$$P(\mathbf{r}|\mathbf{x}) = \prod_{i=1}^N P_i(r_i|\mathbf{x}) \quad (2.3)$$

The  $r_i$ 's might be continuous or discrete. We will illustrate the results on the particular case where  $r_i$  is the number of spikes that cell  $i$  generates during a certain time interval  $\tau$ , with a Poisson statistics (which gives a simple and reasonable description of spikes statistics, see, e.g., Softky and Koch, 1993; Tolhurst et al., 1983) with mean rate  $f_i(\mathbf{x})$ :

$$P_i(r_i|\mathbf{x}) = \frac{(f_i(\mathbf{x}))^{r_i}}{r_i!} \exp(-f_i(\mathbf{x})) \quad (2.4)$$

## 2.2 Mutual Information and Fano bound

Our goal is first to quantify the coding efficiency of the neural population with respect to the classification task at hand, and second to characterize the optimal neural code. Note that we do not address the question of learning or decoding. We thereby do not assume any particular decoding method to extract the estimate  $\hat{\mu}$  nor any particular type of learning process to build the neural representation, so that our approach remains general and applies to the wide class of models that shares the same general assumptions.

Coding efficiency can be quantified making use of information theoretic tools. In the case of population coding of a stimulus in a discrimination task, this has been done for the coding of both continuous (Seung and Sompolinsky, 1993; Brunel and Nadal, 1998) and discrete stimuli (Kang and Sompolinsky, 2001; Kang et al., 2004). In the present context of classification, a relevant quantity is the *mutual information*  $I(\mu, \mathbf{r})$  (to be precisely defined below), that measures the statistical dependency between the two random variables  $\mu$  and  $\mathbf{r}$ . Another natural criterion would be the smallest possible probability  $P_e = P(\hat{\mu} \neq \mu)$  of misclassifying an incoming stimulus. Given a coding scheme  $\mu \rightarrow \mathbf{x} \rightarrow \mathbf{r}$ ,  $P_e$  is the minimal error rate that can be achieved by the best possible estimator,  $\mathbf{r} \rightarrow \hat{\mu}(\mathbf{r})$ . This quantity  $P_e$  is not directly

amenable to analytical treatment. However, as explained below, a standard result in information theory gives an optimal bound on this error probability in term of the mutual information between the neural code and the categories. We will thus base our analysis on this information theoretic criterion.

The mutual information between the categories and the neural activity is defined by (Blahut, 1987):

$$I(\mu, \mathbf{r}) = \sum_{\mu=1}^M q_{\mu} \int d^N \mathbf{r} P(\mathbf{r}|\mu) \ln \frac{P(\mathbf{r}|\mu)}{P(\mathbf{r})} \quad (2.5)$$

where  $P(\mathbf{r})$  is the probability density function (p.d.f.) of  $\mathbf{r}$ :

$$P(\mathbf{r}) = \sum_{\mu=1}^M q_{\mu} P(\mathbf{r}|\mu). \quad (2.6)$$

Here and in all this paper ‘ln’ designates the natural (Neperian) logarithm (information quantities are thus measured in bits if one divides them by  $\ln 2$ ). The mutual information  $I(\mu, \mathbf{r})$  can also be written as

$$I(\mu, \mathbf{r}) = \mathcal{H}(\mu) - \mathcal{H}(\mu|\mathbf{r}) \quad (2.7)$$

where  $\mathcal{H}(\mu) = \mathcal{H}(\{q_{\mu}\}_{\mu=1}^M)$  is the Shannon information (entropy) of the category distribution,

$$\mathcal{H}(\mu) = - \sum_{\mu=1}^M q_{\mu} \ln q_{\mu} \quad (2.8)$$

$\mathcal{H}(\mu)$  corresponds to the information conveyed by the code about the categories in case of a perfect knowledge of the category for any input. The second term  $\mathcal{H}(\mu|\mathbf{r})$  is called the equivocation, and corresponds to the uncertainty in identifying the category knowing the neural response, due to the fluctuations in neural activities and the possibility of having a same input  $\mathbf{x}$  for different categories. It is defined as follows:

$$\mathcal{H}(\mu|\mathbf{r}) = - \int d^N \mathbf{r} P(\mathbf{r}) \sum_{\mu=1}^M Q(\mu|\mathbf{r}) \ln Q(\mu|\mathbf{r}) \quad (2.9)$$

where, according to Bayes’ rule:

$$Q(\mu|\mathbf{r}) = \frac{P(\mathbf{r}|\mu)q_{\mu}}{P(\mathbf{r})} \quad (2.10)$$

Thus, intuitively, minimizing the equivocation  $\mathcal{H}(\mu|\mathbf{r})$  – ie maximizing the mutual information  $I(\mu, \mathbf{r})$  – decreases the probability of error  $P_e$ . This is stated more

rigorously by Fano's optimal inequality (see e.g. Cover and Thomas, 2006, §2.10):

$$\mathcal{H}_b(P_e) + P_e \ln(M - 1) \geq \mathcal{H}(\mu) - I(\mu, \mathbf{r}) \quad (2.11)$$

with  $\mathcal{H}_b(P_e)$  the binary entropy

$$\mathcal{H}_b(P_e) \equiv -P_e \ln P_e - (1 - P_e) \ln(1 - P_e). \quad (2.12)$$

The intuitive meaning of this bound is simple: the left hand side of (2.11) is the information loss due to errors randomly distributed among the  $M$  categories with an error rate  $P_e$ , and the right hand side is the equivocation  $\mathcal{H}(\mu|\mathbf{r})$ , which is the difference between the information that one wants to encode and the one which is actually conveyed. In information theory, a weaker form of the bound, of interest for large values of  $M$ , is commonly used:

$$P_e \geq \frac{\mathcal{H}(\mu|\mathbf{r}) - \ln 2}{\ln(M)} \quad (2.13)$$

Note however that for two categories,  $M = 2$ , the bound (2.11) takes also a simple form:

$$\mathcal{H}_b(P_e) \geq \mathcal{H}(\mu|\mathbf{r}). \quad (2.14)$$

The Fano inequality has already been exploited in the neural network and pattern classification literature (see e.g. Nadal, 1994; Fisher and Principe, 1998; Torkkola and Campbell, 2000). Since the source entropy  $\mathcal{H}(\mu)$  is given, one sees that maximizing the mutual information optimizes the Fano bound.

Making use of the general structure of the code given by equations (2.1) and (2.3), we now consider the large  $N$  behavior of the mutual information (2.5) for a large family of tuning curves.

## 3 Results

### 3.1 The Mutual Information in the Infinite $N$ Limit

Since processing cannot increase information (see e.g. Blahut, 1987, pp. 158-159), the information  $I(\mu, \mathbf{r})$  conveyed by  $\mathbf{r}$  about  $\mu$  is at most equal to the one conveyed by the sensory input  $\mathbf{x}$ :

$$I(\mu, \mathbf{r}) \leq I(\mu, \mathbf{x}) \quad (3.15)$$

Since one has a finite number of categories, the latter is itself at most equal to the entropy  $\mathcal{H}(\mu)$  of the category distribution:

$$I(\mu, \mathbf{x}) \leq \mathcal{H}(\mu) \leq \ln M \quad (3.16)$$

With a model such as (2.2), (2.3), under mild hypotheses on the function  $f$ , and with a distribution of the preferred stimuli covering the full range of possible  $\mathbf{x}$  values, one can see that in the large  $N$  limit the maximal available information is recovered:

$$\lim_{N \rightarrow \infty} I(\mu, \mathbf{r}) = I(\mu, \mathbf{x}) \quad (3.17)$$

In the case of a smooth function  $f$ , this is because for  $N$  large, given  $\mathbf{r}$  the probability of what was the sensory input is sharply peaked around the most probable value  $\mathbf{x}_m(\mathbf{r})$ ; in the limit, every  $\mathbf{r}$  is associated to a single  $\mathbf{x}$  and every  $\mathbf{x}$  to a single  $\mathbf{r}$ : as  $N$  increases the neural code gives a more and more detailed sampling of the  $\mathbf{x}$  distribution. Note that this large  $N$  limit is thus also a high signal-to-noise limit. The asymptotic limit (3.17) can be shown more formally using Laplace/steepest descent method (proceeding as in Brunel and Nadal, 1998). This result is general, and does not assume any optimization of the neural code. We have now to consider  $N$  large but finite, that is the first non trivial correction that depends on the neural code. We discuss two limiting cases: first the general case of smooth tuning curves, and second a simple example of sharp (non smooth) tuning curves. We will see that both cases give qualitatively similar predictions.

## 3.2 Information Content for Large but Finite $N$

### 3.2.1 Smooth Tuning Curves

For smooth enough tuning curves, and under the hypothesis that for any neural activity the maximum likelihood is well defined and unique, we compute (see Appendix A) the mutual information for  $N \gg 1$  and  $K \ll N$ .

In the particular case of a 1-d stimulus (input dimension  $K = 1$ ), we get:

$$I(\mu, x) - I(\mu, \mathbf{r}) = \frac{1}{2} \int dx p(x) \frac{F_{\text{cat}}(x)}{F_{\text{code}}(x)} \quad (3.18)$$

where  $F_{\text{code}}(x)$  is the Fisher information characterizing the sensitivity of  $\mathbf{r}$  with respect to small variations of  $x$ :

$$F_{\text{code}}(x) = - \int d^N \mathbf{r} P(\mathbf{r}|x) \frac{\partial^2 \ln P(\mathbf{r}|x)}{\partial x^2} \quad (3.19)$$

and  $F_{\text{cat}}(x)$  is the Fisher information characterizing the sensitivity of  $\mu$  with respect to small variations of  $x$ :

$$F_{\text{cat}}(x) = - \sum_{\mu=1}^M P(\mu|x) \frac{\partial^2 \ln P(\mu|x)}{\partial x^2} \quad (3.20)$$

which can also be written as

$$F_{\text{cat}}(x) = \sum_{\mu=1}^M \frac{P'(\mu|x)^2}{P(\mu|x)} \quad (3.21)$$

where  $P'(\mu|x) = \partial P(\mu|x)/\partial x$ . Note that  $F_{\text{cat}}$  is independent of the neural code, and that, for  $N$  coding cells,  $F_{\text{code}}$  is of order  $N$ , so that the right hand side of (3.18) is of order  $1/N$  (higher order terms are neglected).

The inverse of the Fisher information is an optimal lower bound on the variance  $\sigma_x^2$  of any unbiased estimator  $\hat{x}(\mathbf{r})$  of  $x$  (Cramer-Rao bound, see e.g. Blahut, 1987):

$$\sigma_x^2 \equiv \int d^N \mathbf{r} P(\mathbf{r}|x) (\hat{x}(\mathbf{r}) - x)^2 \geq \frac{1}{F_{\text{code}}(x)} \quad (3.22)$$

In the more general case of a  $K$ -dimensional sensory input, we get for  $N \gg 1$  and  $K \ll N$  (Appendix A):

$$I(\mu, \mathbf{x}) - I(\mu, \mathbf{r}) = \frac{1}{2} \int d^K \mathbf{x} p(\mathbf{x}) F_{\text{cat}}(\mathbf{x}) : F_{\text{code}}^{-1}(\mathbf{x}) \quad (3.23)$$

where  $F_{\text{code}}(\mathbf{x})$  is the  $K \times K$  Fisher information matrix of the neuronal population:

$$[F_{\text{code}}(\mathbf{x})]_{kl} = - \int d^N \mathbf{r} P(\mathbf{r}|\mathbf{x}) \frac{\partial^2 \ln P(\mathbf{r}|\mathbf{x})}{\partial x_k \partial x_l} \quad (3.24)$$

and  $F_{\text{cat}}(\mathbf{x})$  is the  $K \times K$  Fisher information matrix of the categories:

$$[F_{\text{cat}}(\mathbf{x})]_{kl} = - \sum_{\mu=1}^M P(\mu|\mathbf{x}) \frac{\partial^2 \ln P(\mu|\mathbf{x})}{\partial x_k \partial x_l} \quad (3.25)$$

and ‘:’ stands for the Frobenius inner product on  $K \times K$  real matrices defined as follows: for all  $K \times K$  real matrices  $A, B$ ,

$$A : B \equiv \text{tr}(A^T B) = \sum_{k,l} A_{kl} B_{kl}. \quad (3.26)$$

Equation (3.18) in the  $K = 1$  case, with its generalization (3.23) for an arbitrary input dimension  $K$ , are the main quantitative results of this paper. Considering

the 1-dimensional case for simplicity, one sees that (3.18) contains a nice combination of terms. The Fisher information  $F_{\text{code}}(x)$  is specific to the coding stage  $x \rightarrow \mathbf{r}$ : it tells how well the neural code discriminates nearby sensory inputs. The term  $F_{\text{cat}}(x) = \sum_{\mu=1}^M P'(\mu|x)^2/P(\mu|x)$  is specific to the sensory encoding: it tells whether or not the statistics in the input space are well correlated to the categories.

The optimal code will thus depend directly on the behavior of  $F_{\text{cat}}(x)$ . The most important situation of interest is the one of what we may call “smoothly-overlapping categories”: the identification function  $P(\mu|x)$  has a reasonably smooth S-shape, whose slope  $|P'(\mu|x)|$  is largest near the boundaries between categories (the boundary between two categories  $\mu_1$  and  $\mu_2$  is defined as the location in the input space where  $P(\mu_1|x) - P(\mu_2|x)$  changes sign). There are domains in the input space where  $P(\mu|x)$  is almost zero for every category but one, and domains of transition from one category to another. Examples are given by vowels or colors. This framework corresponds also to experimental setups, such as the one used in Freedman et al. (2003) which involves a continuous set of morphed visual stimuli interpolating between cats and dogs. For such categories, one can already draw some conclusions. First, in domains where the  $P(\mu|x)$  are flat, the Fisher information  $F_{\text{cat}}(x)$  is essentially zero: within such domain the characteristics of the cells (number of cells, width of receptive fields) do not matter, provided the receptive fields cover the full domain. If the number of cell is limited, optimization will lead (in a low-noise limit) to a single cell devoted to each homogeneous domain, with a receptive field (width of the tuning curve) covering the full range of the domain. Note that this also corresponds to the case of non-overlapping and non-abutting categories, for which  $F_{\text{cat}}(x)$  is zero for every  $x$ . Second, this S-shape of the identification function implies that the Fisher information  $F_{\text{cat}}(x)$  is large in the regions where the categories overlap. Note that this is just like the usual Fisher information being the highest where the slope of the tuning curve is the steepest (see Paradiso, 1988; Vogels and Orban, 1990; Seung and Sompolinsky, 1993). If the code is optimized, this must be compensated for by a large value of the Fisher information  $F_{\text{code}}$  in this domain – hence by a higher resolution of the sensory input. As a consequence, we expect more cells to code for the boundary, *ie* their steepest slope will be located in this region. One may then expect two types of optimal codes. What would seem the most natural is to have narrow enough tuning curves, which would therefore give more cells located in the transition region between categories. An alternative is to

have cells broadly tuned, but with an almost flat response within a category and a steep slope in the transition region – in that case the neurons may not have their preferred stimulus within the regions of transition between categories. One should note that the scale, hence ‘narrow’ or ‘broad’, is defined with respect to the width of the transition region between categories. Figure 2 illustrate our main predictions under the hypothesis of narrow tuning curves in the transition region.

One should insist that it is the collective result, as measure by the Fisher information  $F_{\text{code}}$  at the population level, which matters. Hence there can be different ways to achieve a same amount of information, so that it is not obvious to derive fully general statements for individual cells.

It is also interesting to contrast the case of category coding with the one of stimulus coding. In the latter (much studied) case, the coding must perform a ‘density estimation’, having more ressources allocated in proportion to the density  $p(x)$ . In contrast, in the case of category coding, a single cell might be sufficient to code for a homogeneous domain where only one category exists. The ressources have to be allocated to the transition regions, and, as can be seen from formula (3.23), two boundaries having similar  $F_{\text{cat}}$  values will have resources in proportion to the density  $p(x)$  in these regions.

We will now mainly focus the discussion on the interesting part, that is on the coding in the transition regions.

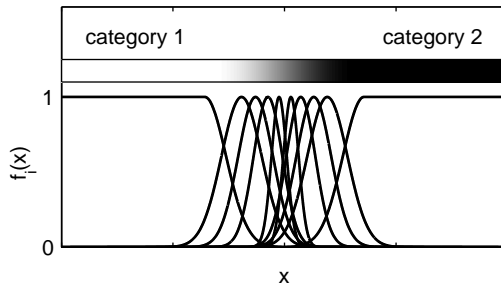


Figure 2: Sketch of an optimal neural code in a simple case involving two smoothly-overlapping categories. The figure shows the tuning curves along the 1-d input space. The gradient map represents the probability  $P(\mu = 1|x)$  of having category 1 knowing the stimulus  $x$ , from value 1 (white) to 0 (black). The stimulus space is divided into three parts: the left part (white), where  $P(\mu = 1|x) = 1$ , the right part (black) where  $P(\mu = 1|x) = 0$ , and a transition region in between.

In the following subsections, we will show that the results above, equations (3.18) and (3.23), are more general, and give some numerical illustrations of optimal codes. Then, in section 3.3, we will also consider a different situation, the one of ill-defined or ill-resolved categories: cases where the sensory input does not contain enough information to clearly associate a domain to a specific class.

### 3.2.2 A Simple Case with Sharp Tuning Curves

To show that the preceding results are more general, we consider a simple illustrative example of sharp tuning curves – that is for which the Fisher information  $F_{\text{code}}$  is infinite for some  $x$  values. We consider a one dimensional sensory input, say  $x$  in  $[0, 1]$ . Each cell has an activity  $r_i$  equal to 1 if  $x$  is in  $[\theta_i, \theta_{i+1}]$ , and 0 otherwise. The width of the  $i$ th tuning curves is thus  $a_i = \theta_{i+1} - \theta_i$ , and we define the preferred stimuli as the centers of the receptive fields,  $x_i \equiv (\theta_i + \theta_{i+1})/2$ .

In this code,  $N$  configurations are possible, each one having one and only one cell activated. We assume that every  $x$  in  $[0, 1]$  may be obtained from at least two categories, and with smooth non constant probabilities  $P(x|\mu)$  as  $x$  varies.

For this particular coding scheme one can easily see that as  $N$  increases the information tends towards  $I(\mu, x)$ . As already seen one has  $I(\mu, \mathbf{r}) \leq I(\mu, x)$ , hence

$$\Delta \equiv I(\mu, x) - I(\mu, \mathbf{r}) \geq 0 \quad (3.27)$$

and this difference  $\Delta$  goes to 0 as  $N$  goes to infinity.

We show in Appendix B that expansion of  $\Delta$  for  $a_i$  small leads to:

$$\Delta = \sum_i \frac{a_i^3}{24} p(x_i) F_{\text{cat}}(x_i) \quad (3.28)$$

where, as before,  $F_{\text{cat}}(x_i) = \sum_{\mu=1}^M P'(\mu|x_i)^2 / P(\mu|x_i)$  and  $P'(\mu|x) = \partial P(\mu|x) / \partial x$ .

Clearly the result (3.28) has the same structure as the one for smooth tuning curves, (3.18). In particular the terms  $\frac{a_i^3}{24}$  should be understood as  $\frac{a_i}{2} \frac{a_i^2}{12}$ , where the first  $a_i$  makes the sum  $\sum_i a_i$  equal to 1, and  $\frac{a_i^2}{12}$  plays the same role as the inverse of the Fisher information  $F_{\text{code}}(x)$ : it gives the (minimal) variance on any estimate of the stimulus given the neural activity. Note however that in the case of smooth tuning curves  $1/F_{\text{code}}(x)$  (hence also  $\Delta$ ) is of order  $1/N$ , whereas here  $a_i^2$  (hence also  $\Delta$ ) is of order  $1/N^2$ .



### 3.2.3 Numerical Illustrations

#### 1-d Numerical Illustration.

*Convergence.* In order to check the domain of validity of formula (3.28) as well as the assumptions we made, we studied a simple example involving two Gaussian categories ( $x^{\mu_1} = 0, x^{\mu_2} = 1, a^{\mu_1} = a^{\mu_2} = 0.25, x \in [0, 1]$ ) and sharply tuned cells. Figure 3 shows a remarkable agreement between the value of  $I(\mu, x) - I(\mu, \mathbf{r})$  and its first correction given above (see Eq. (3.28)), even for low values of  $N$ , in spite of the fact that we assumed  $N \gg 1$  and  $a_i$  small.

We performed the same computation using triangular tuning curves (a shape observed in some empirical data, see e.g. Taube et al., 1990), and again found a notable agreement between  $I(\mu, x) - I(\mu, \mathbf{r})$  and its first correction given by equation (3.18) (figure not shown).

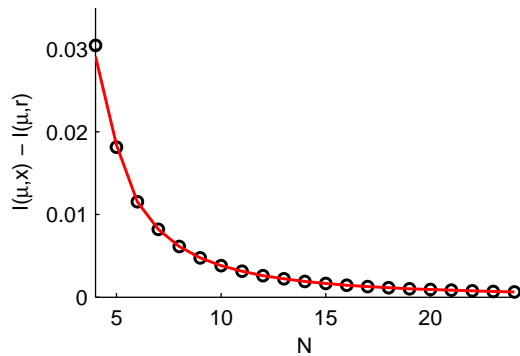


Figure 3: The quantity  $I(\mu, x) - I(\mu, \mathbf{r})$  as a function of the number of neurons  $N$ . The open circles indicates the exact numerical calculation and the line is the second term given by equation (3.28).

*Optimization: Smoothly Tuned Neurons.* Using two Gaussian categories ( $x^{\mu_1} = 0, x^{\mu_2} = 1, a^{\mu_1} = a^{\mu_2} = 0.2, x \in [0, 1]$ ) and  $N = 15$  cells, we performed a numerical minimization of the quantity  $I(\mu, x) - I(\mu, \mathbf{r})$  (given by the right term of equation (3.18)). This optimization over the centers  $\{x_i\}_{i=1}^N$  and the widths  $\{a_i\}_{i=1}^N$  is done using simulated annealing (note that our goal is to find the optimal state, not to mimic any learning mechanism).

The cells have bell-shaped tuning curves (as commonly observed, see e.g. Logothetis et al., 1995):

$$f_i(x) = \exp\left(-\frac{(x - x_i)^2}{2a_i^2}\right) \quad (3.29)$$

and the activity  $r_i$  is given by a Poisson statistics with mean firing rate  $f_i(x)$  (see Eq. (2.4)). Figure 4 shows the tuning curves obtained as well as the Fisher information of the corresponding population. In compliance with our intuitive analysis, we find more cells between categories with sharper tuning curves.

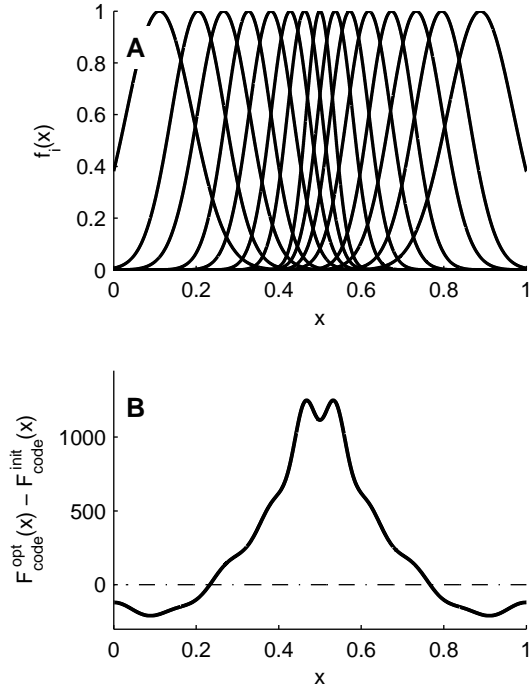


Figure 4: (A) Optimal tuning curves obtained numerically. (B) As a function of  $x$ , difference between the Fisher information  $F_{\text{code}}(x)$  for the optimal code, and the one for the equidistributed distribution of preferred stimuli.

*Optimization: Sharply Tuned Neurons.* Using the same categories as before and  $N = 15$  cells, we performed a minimization of the quantity  $I(\mu, x) - I(\mu, \mathbf{r})$  using a gradient descent on the right term of equation (3.28). Figure 5 shows the final configuration we obtained, in agreement with our intuitive analysis.

## 2-d Numerical Illustration.

In a two-dimensional case involving two categories, we would expect (1) cells to be more sharply tuned along the most relevant dimension, and (2) within the relevant dimension, cells to be more sharply tuned in the regions at the boundary than in other regions.

In order to investigate these predictions, we studied a 2-d numerical example involv-

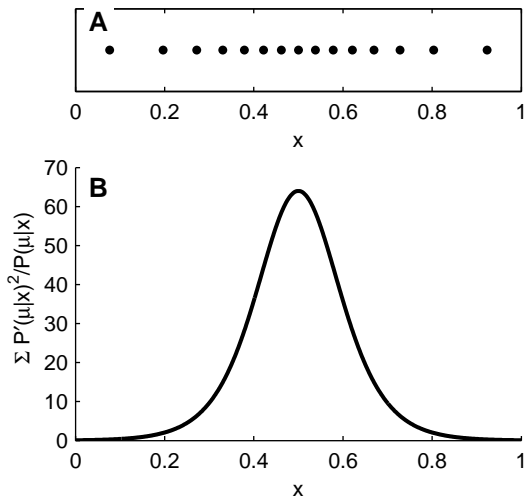


Figure 5: (A) Optimal configuration of the  $\{x_i, i = 1, \dots, N\}$ , for  $N = 15$ . (B) The Fisher information  $F_{\text{cat}}(x)$  as function of  $x$ .

ing two Gaussian categories with parameters:

$$\mathbf{x}^{\mu_1} = \begin{pmatrix} -2 \\ 0 \end{pmatrix}, \mathbf{x}^{\mu_2} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \mathbf{a}^{\mu_1} = \mathbf{a}^{\mu_2} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

so that the  $x$ -dimension is more relevant for the categorization task than the  $y$ -dimension. Initially,  $N = 63$  ( $N = N_x \times N_y = 9 \times 7$ ) cells are equidistributed over the domain  $\{-5 \leq x \leq 5, -3 \leq y \leq 3\}$ . The neurons have Gaussian tuning curve  $f_i(x, y)$  (with a width initially equal to  $5/(N_x - 1)$  along both dimension) and their activity  $r_i$  is given by a Poisson statistics with mean firing rate  $f_i(x, y)$  (see Eq. (2.4)). Figure 6 shows the configuration obtained from numerical optimization, in agreement with the predictions made above: more cells code for the region where categories overlap. One should note that the pictured configuration is suboptimal, due to the chosen algorithmic procedure which is based on local changes with tuning curves restricted to 2d Gaussian shapes. As discussed in section 3.2.1 and emphasized by figure 2, a large single receptive field covering the contiguous domains where  $F_{\text{cat}}(x)$  is almost zero would be sufficient in order to maximize information.

### 3.3 Ill-resolved Categories

Let us now briefly consider cases to be contrasted with the preceding ones, namely situations where the categories are ill resolved in the sensory space. This includes

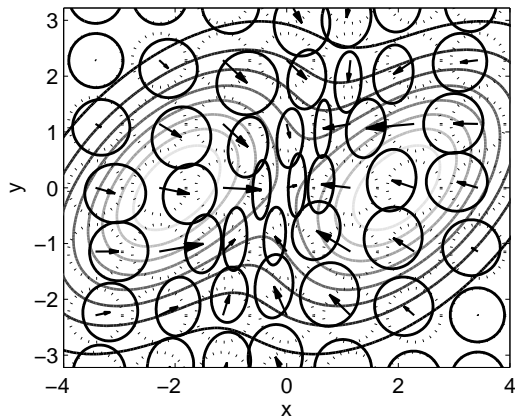


Figure 6: The gray dashed contour represents the density distribution  $P(\mathbf{x})$  of the stimuli  $\mathbf{x} = (x, y)$ , the two inner contours corresponding to the center of the two categories. The dashed circles represent the initial cells, and the solid ellipses depict the optimal configuration obtained numerically. An arrow connects the initial center of a cell to its optimal one. The length of a given axis of an ellipse is proportional to the tuning curve width of the corresponding cell along this dimension.

cases of high noise, single-spike processing, or categories for which the sensors are not well adapted.

In such cases where information on what has to be encoded is low, it is clear that optimal coding will be qualitatively different than in cases of high signal-to-noise ratio. This has been shown explicitly for the coding of a continuous quantity (Stein, 1967; Rieke et al., 1997; Brunel and Nadal, 1998; Butts and Goldman, 2006). There, stimulus-specific cells cannot do better than (crudely) coding for their preferred stimuli (Butts and Goldman, 2006), and the optimal tuning curves are then of all-or-none type – maximal rate in some range around the preferred stimulus, and minimal rate outside (Brunel and Nadal, 1998). In the present context, we expect that no useful information can be extracted from class boundaries, and thus to observe opposite properties compared to the previous cases: adaption to such noisy-like regimes should make neural cells flow away from the class boundaries. We show below that is indeed the case, by considering two qualitatively different examples: the one of two categories with strong overlap in the input space, and the very short-time limit (single spike processing).

### 3.3.1 Ill-defined Classes

We first consider a case for which the asymptotic behavior of the mutual information is still given by the large  $N$  and large time regime (formula (3.18) or (3.28)), but with a structure of  $P(\mu|x)$  that does not give a clear enough separation between categories. By this we mean that the Fisher information  $F_{\text{cat}}(x)$  remains small in all the transition region.

A specific example is obtained by taking  $M = 2$  categories,  $K = 1$ , with  $x \in [0, 1]$ ,  $q_1 = q_2 = 1/2$ ,  $P(x|\mu = 1) = 2x$ ,  $P(x|\mu = 2) = 2(1 - x)$ , so that  $p(x)$  is the uniform distribution on  $[0, 1]$ , and  $P(\mu = 1|x) = x$ ,  $P(\mu = 2|x) = 1 - x$ . Figure 7 gives the plot of  $F_{\text{cat}}(x)$  as function of  $x$ , and the optimal distribution of preferred stimuli numerically computed for this example. One sees that the Fisher information  $F_{\text{cat}}(x)$  is rather flat with a minimum in the middle, leading as expected to less cells at this location.

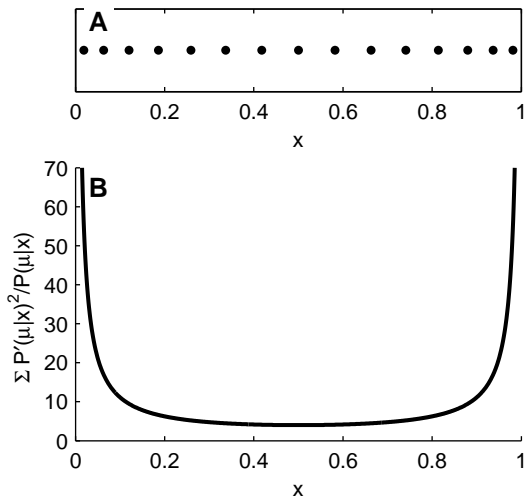


Figure 7: (A) Optimal configuration of the  $\{x_i, i = 1, \dots, N\}$ , for  $N = 15$ . (B) The Fisher information  $F_{\text{cat}}(x)$  as function of  $x$ .

This particular example can be seen as a counter-example to the general results of section 3.2.1, showing the importance of the hypothesis of an S-shape for the behavior of  $F_{\text{cat}}(x)$  at the category boundaries.

### 3.3.2 Short Time Limit

We turn now to the case of very short duration processing, for which one cannot even use the result (3.18) for the mutual information. As in Brunel and Nadal (1998),

we consider the limit where at most one cell emits one spike. Explicitly, we assume that one looks at the neural activity during a short time window  $[0, \tau]$  so that

$$\begin{aligned} P_i(r_i = 1|\mathbf{x}) &= f_i(\mathbf{x}) \tau \\ P_i(r_i = 0|\mathbf{x}) &= 1 - f_i(\mathbf{x}) \tau \end{aligned} \quad (3.30)$$

with  $0 < f_{min} \leq f_i(\mathbf{x}) \leq f_{max}$  for every input  $\mathbf{x}$  and every cell  $i$ . Then the mutual information  $I(\mu, \mathbf{r})$  can be directly computed at first order in  $\tau$ , that is assuming  $\tau N f_{max} \ll 1$ . One gets

$$I(\mu, \mathbf{r}) = \tau \sum_i \sum_{\mu=1}^M q_\mu \bar{f}_i^\mu \ln \frac{\bar{f}_i^\mu}{\bar{f}_i} \quad (3.31)$$

where  $\bar{f}_i = \int d^K \mathbf{x} P(\mathbf{x}) f_i(\mathbf{x})$  is the mean rate of cell  $i$  averaged over all possible inputs, and  $\bar{f}_i^\mu = \int d^K \mathbf{x} P(\mathbf{x}|\mu) f_i(\mathbf{x})$  its mean rate conditional to the category  $\mu$ . Since  $\sum_{\mu=1}^M q_\mu \bar{f}_i^\mu = \bar{f}_i$ , defining  $q_{i,\mu} \equiv q_\mu \bar{f}_i^\mu / \bar{f}_i$  which measures the probability that  $i$  fires given the category  $\mu$  relative to its mean global firing rate, one may write

$$I(\mu, \mathbf{r}) = \tau \sum_i \bar{f}_i \sum_{\mu=1}^M q_{i,\mu} \ln \frac{q_{i,\mu}}{q_\mu} \quad (3.32)$$

One sees that the mutual information is maximized by maximizing the Kullback divergence between the probabilities  $\{q_{i,\mu}, \mu = 1, \dots, M\}$  and  $\{q_\mu, \mu = 1, \dots, M\}$ : each cell must be as specific as possible to one and only one category. This implies that, in an optimized code, most cells will have a receptive field avoiding any domain in the input space where categories overlap. A numerical illustration, involving the two Gaussian categories of section 3.2.2, is presented on figure 8.

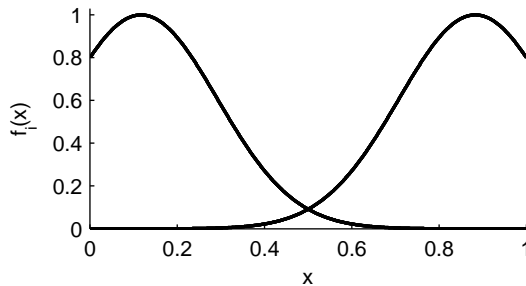


Figure 8: Optimal tuning curves obtained numerically, for  $N = 10$  cells with bell-shaped receptive field. Half of the cells overlap on the right side of the boundary, the other half on the left side.

## 4 Discussion

### 4.1 Perceptual Consequences

We now discuss the consequences of the main quantitative results presented section 3.2. For the sake of simplicity, we will essentially develop the one-dimensional case. As we have seen in section 3.2.1, the category related Fisher information  $F_{\text{cat}}(x) = \sum_{\mu=1}^M P'(\mu|x)^2/P(\mu|x)$  is typically the greatest in the regions near the boundaries between categories. Thus, if the number  $N$  of neurons is limited, maximizing the mutual information between the categories  $\mu$  and the neural activity  $\mathbf{r}$  entails that the Fisher information  $F_{\text{code}}(x)$  is greater at the boundaries between categories (see, e.g., figure 4). This Fisher information is related to the discriminability  $d'$  (see e.g. Green and Swets, 1988) of two stimuli  $x$  and  $x + \delta x$  according to :  $d' = |\delta x| \sqrt{F_{\text{code}}(x)}$  (Seung and Sompolinsky, 1993). As a result, the higher  $F_{\text{code}}(x)$ , the more discriminable two sensory inputs  $x$  and  $x + \delta x$  in the perceptual space given by the output of the neuronal population.

In other words, maximizing the mutual information between the categories  $\mu$  and the neural activity  $\mathbf{r}$  – which any learning mechanism that aims at minimizing the probability of error should do – has the consequence of altering our perception. More precisely, the discriminability will be greater between categories than within, a perceptual phenomena traditionally called *categorical perception* (Harnad, 1987). Categorical perception was first described within the field of speech perception as an innate process specific to human speech that implied high discriminability between items from different (phonemic) categories and zero discriminability within a category (Liberman et al., 1957). This strong version of categorical perception was subsequently undermined: not only this phenomenon can be acquired (Abramson and Lisker, 1970) but it is also not specific to speech (Goldstone, 1994) nor to human (Kuhl and Padden, 1983). Our result fits well into this framework, for it can apply to any modality, might be induced by learning, and does not assume anything specific to human. Moreover, it also indicates that categorical perception is not a mere by-product of category learning but serves a function, that is to minimize classification errors. Note however that we do not claim that other processes, such as top-down influences (see e.g. Sugase et al., 1999; Li et al., 2004), memory effects, or labeling, might not alter perception. Our results give an optimal bound on discriminability based on a purely sensory driven mechanism, and thus give credit to

a sensory account for categorical perception.

Another way to present the effects induced by the change in the neural configuration is in terms of compression/expansion of the perceptual space defined by the output of the neuronal population. If discriminability is higher (respectively lower) after learning than before, then the perceptual space can be seen as expanded (resp. contracted). Thus, whether category learning induce within-category compression and/or between-category expansion (Goldstone, 1994; Livingston et al., 1998) might depend on the initial ability of the neuronal population. In the numerical example pictured by figure 4, we see that, compared to the uniform initial configuration of preferred stimuli, there is acquired distinctiveness at the class boundary and acquired similarity within classes. Such a warping of the perceptual space is related to a phenomenon found in speech perception literature called the *perceptual magnet effect* (Kuhl, 1991; Guenther et al., 1999), stating that discriminability is lower around prototypical stimuli than around non-prototypical ones. Those perceptual phenomena will be more extensively studied in a forthcoming paper (Bonnasse-Gahot, in preparation).

## 4.2 Categorization and the Inferotemporal Cortex

The inferotemporal cortex (ITC) in the monkey has been identified as a site for object recognition and identification (see Tanaka, 1996, for a review), and object categorization (see e.g. Hung et al., 2005; Kiani et al., 2007). Many studies suggest that population coding is a strategy widely used in this area (see, e.g., Young and Yamane, 1992; Logothetis et al., 1995; Vogels, 1999). In addition, some experiments have brought to light the influence of visual experience on the properties of cells. For instance, Kobatake et al. (1998) have shown that perceptual learning on a discrimination task leads to an increase in the proportion of cells in the ITC responsive to some of the training stimuli. As already mentioned, Sigala and Logothetis (2002) have shown that categorization training increases sensitivity of IT neurons to the features relevant for the categorization at hand. Thus, those results makes the ITC a perfect candidate for an empirical evaluation of our predictions.

Accordingly, Thomas et al. (2001) have shown (1) that from the recordings of a population of inferotemporal neurons, the categories can be recovered making use of an unsupervised learning algorithm applied to these data, and (2) that the most informative neurons about the category membership are those responding to both



categories, which is consistent with our analysis. As we have seen, a consequence of category learning is an expansion of the perceptual space between categories and a contraction within. This leads to neural responses that clusters in the output space according to the category the stimuli belong to. This phenomenon, that generates categorical perception, finds support in the study by Wilson and DeBauche (1981) showing that monkeys with inferotemporal lesions do not exhibit categorical perception, and suggesting that the output of the ITC is the “effective stimulus”, in the sense – which corresponds to our point of view – of being the representation which defines the perceptual space. In line with this statement, Kiani et al. (2007) have shown that there is a significant correlation between the behavioral ability of discriminating between pairs of stimuli and the distance between neural IT responses. However, some studies do not find any impact of categorization training on the inferotemporal cells properties, as the one by Op de Beeck et al. (2001). Interestingly, Op de Beeck et al. (2001) have also failed to find any modification on the metric of the perceptual space, which is consistent with our analysis section 4.1. Moreover, although Freedman et al. (2003) conclude from their study on the influences of category learning in the ITC and prefrontal cortex that category learning implies (almost) no change in the neural properties of the ITC, a closer look to their data shows that almost half of all recorded inferotemporal neurons have preferred stimuli located at the class boundary (Knoblich et al., 2002, Fig. 12). Since in addition, according to Knoblich et al. (2002), the tuning is narrow, this empirical result is qualitatively in agreement with our predictions.

Finally, in the recent work of Koida and Komatsu (2007) recording color-selective neurons in the ITC, the distribution of preferred stimuli is found to be more concentrated at the center of the two categories that were investigated (the two colors ‘red’ and ‘green’). However, as the tuning is broad, the stimulus-discriminating part of the tuning curves do fall within the boundary region, in compliance with our results. Interestingly, this experimental study also tackles the issue of the task-dependent modulation of the activity of a neural map. A same population of cells is shown to be used for both categorization and discrimination. It is found that during a categorization task, the neuronal population is modulated so as to detect changes from one category to another. In line with our predictions, this is obtained through enhancing neural activities of cells with the steepest part of their tuning curve located in the transition region, and decreasing the neural activities of the others (see figures 2 and 6 in Koida and Komatsu, 2007).

### 4.3 Conclusion and Future Work

A large part of experimental works on neural selectivity characterize this selectivity by the cells preferred stimuli only. There is, however, an increasing number of both experimental (see e.g. Vogels and Orban, 1990; Han et al., 2007) and theoretical works (including the present one), showing that a more detailed description of the tuning curves is mandatory in order to assess the actual role of a cell or of a population of cells. In the case of categorization, extrapolating from a study on the auditory cortex in the rat, Han et al. (2007) anticipate that “tuning curve slopes [should be] aligned at the location of the categorical boundaries”. In the present paper we have proposed an information theoretic approach which essentially support this statement, and allows more generally to characterize the efficiency of a neural code in a classification task.

Our main quantitative results concern the case of smoothly overlapping categories, for which we have exhibited a formula, Eq. (3.23), giving the mutual information between the activity of a neuronal population and a set of discrete categories in the limit of a large (but finite) number of coding cells. This formula takes the simple form of an average (over the input space) of the ratio between two Fisher-information values: in the denominator, the (usual) Fisher information  $F_{\text{code}}(x)$  giving the sensitivity of the population code to small changes in the input space; and in the numerator, the category related Fisher information  $F_{\text{cat}}(x)$  given in terms of the posterior probability of each class,  $\{P(\mu|x), \mu = 1, \dots, M\}$ .

As discussed in section 3, this category related Fisher information  $F_{\text{cat}}(x)$  is typically the greatest at the boundaries between classes. As the number of neurons is limited, category learning implies (usual) Fisher information  $F_{\text{code}}$  to be the greatest in these regions, result that gives a precise formulation and explanation of categorical perception (see section 4.1). This, and other perceptual phenomena, will be more extensively discussed in a forthcoming paper (Bonnasse-Gahot, in preparation).

Moreover, considering the optimal configuration of a population of neurons in terms of the preferred stimuli of the cells and the width of their receptive field, our results predict that, if there is indeed optimization, (1) cells will be specifically tuned to the dimensions relevant for the classification task at hand; (2) more neurons will be allocated at the class boundaries, meaning that (3) more steep slopes should be found in the transition regions between categories. All these predictions find support in recent neurophysiological experiments (see section 4.2), but clearly future research

is needed to quantitatively check all our predictions.

The above main results have been obtained in the case of well-defined, smoothly overlapping, classes and sufficiently large signal-to-noise ratio. One should remind that, as shown section 3.1, optimization of the neural code is necessary to increase information content only under the constraint of limited resources – otherwise, in the limit of a very large number of cells, even a random code may convey all the (finite) information about the categories. We have also studied ill-defined classes and the short-time limit (single-spike processing): in these “noisy” cases, the main result is that, contrary to the previous situation, the cells will avoid the class boundaries. More generally, these limiting cases indicate how the general conclusion on the sharpness of tuning curves at the class boundaries is modulated in the presence of any kind of noise (sensory or synaptic noise, small number of cells, short-duration processing...): as less information can be extracted, the tuning becomes broader, with less steep slope at the class boundaries.

Several directions for future work might be considered. First, an important hypothesis in our calculation is that, given the input, the activities of the neurons are statistically independent. Recent work has already shown that taking into account correlations might change the coding efficiency of a neuronal population (see Averbeck et al., 2006, for a review). In the theoretical analysis of the efficiency of population coding in the case of discrimination tasks, the specific role of correlations has been the subject of several papers (Abbott and Dayan, 1999; Yoon and Sompolinsky, 1999; Sompolinsky et al., 2001; Seriès et al., 2004). These studies have shown that different scenarios may lead to opposite effects. In some case the redundancy in the neural activity limits the total amount of available information. For strong correlations, this may lead, in the large  $N$  limit, to an encoding which is essentially equivalent to have an effective finite number of cells (Sompolinsky et al., 2001) – in such a case, even an infinite number of cells may not allow to recover the full information that the stimulus convey about the categories. In contrast, some types of correlations can lead to more efficient coding, with a qualitative difference concerning the shape of optimal tuning curves (Seriès et al., 2004) – a rather broad tuning becoming optimal. One can thus expect that, in order to have a given amount of information conveyed by the neural population about the categories, compared to a coding with uncorrelated cells, the number of cells coding for the transition region between two categories should be smaller (respectively larger), and the slopes of

their tuning curves should be less steep (resp. steepest) for correlations increasing (resp. decreasing) the information conveyed about the stimulus. A necessary extension will therefore be to study the generalization of our analysis to cases involving correlations between cells. In particular, it is not clear under which conditions the analytical expression (3.23) of the mutual information remains valid.

We have discussed, section 2.1, the experimental and computational motivations for considering a low-dimensional input space. One may add that, in smooth enough cases, the boundary between two categories, defined by the locus of  $P(\mu = 1|\mathbf{x}) = P(\mu = 2|\mathbf{x})$ , is an hypersurface in the  $\mathbf{x}$ -space. The most discriminant dimension is then, locally, the normal to this surface. The 1-d case corresponds thus to a simplification of this general picture where one is looking only along this most discriminant dimension near one particular point of the boundary. One can note also that the calculations we have made do not assume that the input dimension  $K$  is small, but, more exactly, that it is small compared to the (very large) number of cells  $N$ . Yet, it would be interesting to consider  $K$  of order  $N$  – one may expect the case  $K = \alpha N$ , with  $\alpha$  given and  $N \rightarrow \infty$ , to be amenable to analytical treatment.

Clearly specific systems should be studied in more depth, notably in order to investigate more specific input space than the idealized  $K$ -dimensional static space we posited. For instance, in the case of vowels, although the corresponding space might be seen as a low-dimensional space given by the first formants, the physical signal itself not only lies in a high-dimensional space but also presents a temporal component.

The recent experimental work by Kiani et al. (2007) do not only show that responses of a population of inferotemporal neurons represent categories, but also that they represent *relationship* between categories. However, the information conveyed by a neural code on a discrete set of categories does not say anything about the *structure* of these categories. In the extreme case of non-abutting categories, shortly evoked in this paper, the code maximizing mutual information is obviously to have, in a low noise limit, one cell per class, each one having its receptive field covering the input space specific to a same class. On the contrary, taking into account relationship – whatever is the similarity criterion – between categories might be achieved thanks to cells spanning several similar categories. Extending the information theoretic approach in order to cope with the structure of categories, such as a hierarchical organization, constitutes an exciting challenge.

As mentioned in section 2.1, we do not assume any particular type of learning or

decoding method. As a consequence, our result remains general and applies to any model that shares the same general assumptions. Another interesting direction therefore consists in characterizing learning mechanisms that aim at attaining an optimal code, for both supervised and unsupervised learning of categories, and to explore plausible decoding schemes. In the case of a continuous stimulus, Pouget et al. (1998) have shown that maximum likelihood decoding can be performed with an recurrent neural network. It would be benefiting to see if a similar strategy can be used in the case of categorization.

Finally, it would be interesting to explore the relevance of our approach to the analysis of motor maps that codes for movements (see, e.g., Georgopoulos et al., 1986), in cases where prototypes of categories have to be produced, as in speech (production of phonemes or words).

## Acknowledgments

This work is part of a project “Acqlang” supported by the French National Research Agency (ANR-05-BLAN-0065-01). LBG acknowledges a fellowship from the DGA. JPN is a CNRS member.

The initial motivation for this work comes from (psycho- and neuro-) computational issues in the perception of phonemes: we thank Sharon Peperkamp and Janet Pierrehumbert for introducing us to this topic and for valuable discussions. LBG is grateful to the members of the Laboratoire de Sciences Cognitives et Psycholinguistique de l’ENS, especially to Emmanuel Dupoux, for numerous and stimulating discussions. We acknowledge useful inputs from the referees, and most especially, we thank one of them for a detailed list of constructive comments.

## A Derivation of Equation (3.23)

The goal of this appendix and the following one is to derive equations (3.18), (3.23) and (3.28). Remark: given the simplicity and the underlying identity of these results, we do not expect our derivations to be the simplest ones.

When  $N$  goes to  $\infty$ , we expect the mutual information  $I(\mu, \mathbf{r})$  to converge towards  $I(\mu, \mathbf{x})$ , and we are interested in the first non trivial correction to this asymptotic limit. We thus compute for large  $N$  the difference

$$\Delta \equiv I(\mu, \mathbf{x}) - I(\mu, \mathbf{r}) \geq 0. \quad (\text{A.1})$$

One can write

$$\Delta = - \int d^K \mathbf{x} d^N \mathbf{r} P(\mathbf{r}|\mathbf{x}) \phi(\mathbf{x}|\mathbf{r}) \quad (\text{A.2})$$

where

$$\phi(\mathbf{x}|\mathbf{r}) \equiv \sum_{\mu=1}^M p(\mathbf{x}) P(\mu|\mathbf{x}) \ln \frac{P(\mu|\mathbf{r})}{P(\mu|\mathbf{x})} \quad (\text{A.3})$$

We follow the same approach as in Brunel and Nadal (1998). The first step consists in integrating over  $\mathbf{x}$ . Taking the large  $N$  limit, we show that the leading order of the right term of equation (A.2) is zero. We then seek for the first correction using Laplace/steepest descent method. The last step eventually consists in integrating over  $\mathbf{r}$ .

We introduce  $G(\mathbf{r}|\mathbf{x})$  defined as :

$$G(\mathbf{r}|\mathbf{x}) \equiv \frac{1}{N} \ln P(\mathbf{r}|\mathbf{x}) \quad (\text{A.4})$$

and assume that it has a single global maximum at  $\mathbf{x} = \mathbf{x}_m(\mathbf{r})$ . We can rewrite equation (A.2) in the following way:

$$\Delta = - \int d^K \mathbf{x} d^N \mathbf{r} e^{NG(\mathbf{r}|\mathbf{x})} \phi(\mathbf{x}|\mathbf{r}) \quad (\text{A.5})$$

**Integration over  $\mathbf{x}$ .** In order to integrate equation (A.2) over  $x$ , let us first show that

$$\phi(\mathbf{x}_m(\mathbf{r})|\mathbf{r}) = 0 \quad (\text{A.6})$$

We begin by evaluating

$$P(\mu|\mathbf{r}) - P(\mu|\mathbf{x}_m(\mathbf{r})) = \frac{P(\mathbf{r}|\mu)q_\mu}{P(\mathbf{r})} - P(\mu|\mathbf{x}_m(\mathbf{r})) \quad (\text{A.7})$$

$$= \frac{1}{P(\mathbf{r})} \int d^K \mathbf{x} \exp(NG(\mathbf{r}|\mathbf{x})) \varphi_\mu(\mathbf{x}|\mathbf{r}) \quad (\text{A.8})$$

with

$$\varphi_\mu(\mathbf{x}|\mathbf{r}) = p(\mathbf{x}) [P(\mu|\mathbf{x}) - P(\mu|\mathbf{x}_m(\mathbf{r}))] \quad (\text{A.9})$$

By using saddle-point method and assuming that  $N \gg 1$  and  $K \ll N$ , we find :

$$P(\mu|\mathbf{r}) - P(\mu|\mathbf{x}_m(\mathbf{r})) = \frac{1}{P(\mathbf{r})} \exp(NG_m(\mathbf{r})) \sqrt{\frac{(2\pi)^K}{\det \mathbb{H}(-NG(\mathbf{r}|\mathbf{x}))|_{\mathbf{x}_m(\mathbf{r})}}} \\ \times \frac{1}{2} \text{tr} \left( \mathbb{H}(\varphi_\mu(\mathbf{x}|\mathbf{r}))|_{\mathbf{x}_m(\mathbf{r})} \mathbb{H}(-NG(\mathbf{r}|\mathbf{x}))^{-1}|_{\mathbf{x}_m(\mathbf{r})} \right) \quad (\text{A.10})$$

where  $\mathbb{H}(\xi(\mathbf{x}))|_{\mathbf{x}_m(\mathbf{r})}$  is the hessian matrix of  $\xi$ ,

$$\mathbb{H}_{kl}(\xi(\mathbf{x})) = \frac{\partial^2 \xi(\mathbf{x})}{\partial x_k \partial x_l} \quad (\text{A.11})$$

evaluated at  $\mathbf{x}_m(\mathbf{r})$ .

Since

$$P(\mathbf{r}) = p(\mathbf{x}_m(\mathbf{r})) \exp(NG_m(\mathbf{r})) \sqrt{\frac{(2\pi)^K}{\det \mathbb{H}(-NG(\mathbf{r}|\mathbf{x}))|_{\mathbf{x}_m(\mathbf{r})}}} \quad (\text{A.12})$$

we have

$$P(\mu|\mathbf{r}) - P(\mu|\mathbf{x}_m(\mathbf{r})) = \frac{1}{2p(\mathbf{x}_m(\mathbf{r}))} \text{tr} \left( \mathbb{H}(\varphi_\mu(\mathbf{x}|\mathbf{r}))|_{\mathbf{x}_m(\mathbf{r})} \mathbb{H}(-NG(\mathbf{r}|\mathbf{x}))^{-1}|_{\mathbf{x}_m(\mathbf{r})} \right) \quad (\text{A.13})$$

As a result,

$$\phi(\mathbf{x}_m(\mathbf{r})|\mathbf{r}) \equiv \sum_{\mu=1}^M p(\mathbf{x}_m(\mathbf{r})) P(\mu|\mathbf{x}_m(\mathbf{r})) \ln \frac{P(\mu|\mathbf{r})}{P(\mu|\mathbf{x}_m(\mathbf{r}))} \quad (\text{A.14})$$

$$= \sum_{\mu=1}^M p(\mathbf{x}_m(\mathbf{r})) P(\mu|\mathbf{x}_m(\mathbf{r})) \ln \left( 1 + \frac{P(\mu|\mathbf{r}) - P(\mu|\mathbf{x}_m(\mathbf{r}))}{P(\mu|\mathbf{x}_m(\mathbf{r}))} \right) \quad (\text{A.15})$$

According to equation (A.13),  $P(\mu|\mathbf{r}) - P(\mu|\mathbf{x}_m(\mathbf{r}))$  is of order  $1/N$ , which entails, as  $\ln(1+z) \approx z$  when  $z \ll 1$ , that

$$\phi(\mathbf{x}_m(\mathbf{r})|\mathbf{r}) = \frac{1}{2} \sum_{\mu=1}^M \text{tr} \left( \mathbb{H}(\varphi_\mu(\mathbf{x}))|_{\mathbf{x}_m(\mathbf{r})} \mathbb{H}(-NG(\mathbf{r}|\mathbf{x}))^{-1}|_{\mathbf{x}_m(\mathbf{r})} \right) \quad (\text{A.16})$$

$$= \frac{1}{2} \text{tr} \left( \mathbb{H} \left( \sum_{\mu=1}^M \varphi_\mu(\mathbf{x}) \right) |_{\mathbf{x}_m(\mathbf{r})} \mathbb{H}(-NG(\mathbf{r}|\mathbf{x}))^{-1}|_{\mathbf{x}_m(\mathbf{r})} \right) \quad (\text{A.17})$$

Now, as  $\sum_{\mu=1}^M \varphi_{\mu}(\mathbf{x}) = 0$  we have demonstrated that  $\phi(\mathbf{x}_m(\mathbf{r})|\mathbf{r}) = 0$ .

We can return to equation (A.2), and apply saddle-point method knowing that  $\phi(\mathbf{x}_m(\mathbf{r})|\mathbf{r}) = 0$ . This leads to :

$$\begin{aligned} \Delta = & - \int d^N \mathbf{r} \exp (NG_m(\mathbf{r})) \sqrt{\frac{(2\pi)^K}{\det \mathbb{H}(-NG(\mathbf{r}|\mathbf{x}))|_{\mathbf{x}_m(\mathbf{r})}}} \\ & \times \frac{1}{2} \text{tr} \left( \mathbb{H}(\phi(\mathbf{x})|\mathbf{r})|_{\mathbf{x}_m(\mathbf{r})} \mathbb{H}(-NG(\mathbf{r}|\mathbf{x}))^{-1}|_{\mathbf{x}_m(\mathbf{r})} \right) \end{aligned} \quad (\text{A.18})$$

Recalling that  $P(\mu|\mathbf{r}) - P(\mu|\mathbf{x}_m(\mathbf{r})) \sim O(\frac{1}{N})$ , it is straightforward to show that :

$$\frac{1}{p(\mathbf{x}_m(\mathbf{r}))} \frac{\partial \phi(\mathbf{x}|\mathbf{r})}{\partial x_k \partial x_l} \Big|_{\mathbf{x}_m(\mathbf{r})} = - \sum_{\mu=1}^M \frac{1}{P(\mu|\mathbf{x}_m(\mathbf{r}))} \frac{\partial P(\mu|\mathbf{x})}{\partial x_k} \Big|_{\mathbf{x}_m(\mathbf{r})} \frac{\partial P(\mu|\mathbf{x})}{\partial x_l} \Big|_{\mathbf{x}_m(\mathbf{r})} \quad (\text{A.19})$$

ie

$$\frac{1}{p(\mathbf{x}_m(\mathbf{r}))} \mathbb{H}(\phi(\mathbf{x}|\mathbf{r}))|_{\mathbf{x}_m(\mathbf{r})} = - \sum_{\mu=1}^M \frac{1}{P(\mu|\mathbf{x}_m(\mathbf{r}))} \nabla P(\mu|\mathbf{x})|_{\mathbf{x}_m(\mathbf{r})} \nabla P(\mu|\mathbf{x})|_{\mathbf{x}_m(\mathbf{r})}^{\top} \quad (\text{A.20})$$

where  $\nabla \xi(\mathbf{x})|_{\mathbf{x}_m(\mathbf{r})}$  is the column vector of the partial derivatives of  $\xi$

$$\nabla \xi(\mathbf{x}) = (\dots, \partial \xi(\mathbf{x})/\partial x_k, \dots)^{\top} \quad (\text{A.21})$$

evaluated at  $\mathbf{x}_m(\mathbf{r})$ .

Putting equations (A.12), (A.18) and (A.20) together eventually leads to:

$$\Delta = \frac{1}{2} \int d^N \mathbf{r} P(\mathbf{r}) \left( \sum_{\mu=1}^M \frac{1}{P(\mu|\mathbf{x})} \nabla P(\mu|\mathbf{x}) \nabla P(\mu|\mathbf{x})^{\top} \right) \Big|_{\mathbf{x}_m(\mathbf{r})} : \mathbb{H}(-NG(\mathbf{r}|\mathbf{x}))^{-1} \Big|_{\mathbf{x}_m(\mathbf{r})} \quad (\text{A.22})$$

where ‘:’ denotes the Frobenius inner product on  $\mathcal{M}_K(\mathbb{R})$ , defined as follows:

$$\forall A, B \in \mathcal{M}_K(\mathbb{R}), \quad A : B = \text{tr}(A^{\top} B) = \sum_{k,l} A_{kl} B_{kl}. \quad (\text{A.23})$$

**Integration over  $\mathbf{r}$ .** By proceeding as Brunel and Nadal (1998, pp.1753-1754) in order to integrate over  $\mathbf{r}$ , we get:

$$I(\mu, \mathbf{x}) - I(\mu, \mathbf{r}) = \frac{1}{2} \int d^K \mathbf{x} p(\mathbf{x}) \left[ \sum_{\mu=1}^M \frac{1}{P(\mu|\mathbf{x})} \nabla P(\mu|\mathbf{x}) \nabla P(\mu|\mathbf{x})^{\top} \right] : F_{\text{code}}^{-1}(\mathbf{x}) \quad (\text{A.24})$$



where  $F_{\text{code}}(\mathbf{x})$  is the  $K \times K$  Fisher information matrix of the neuronal population,

$$[F_{\text{code}}(\mathbf{x})]_{kl} = - \int d^N \mathbf{r} P(\mathbf{r}|\mathbf{x}) \frac{\partial^2 \ln P(\mathbf{r}|\mathbf{x})}{\partial x_k \partial x_l} \quad (\text{A.25})$$

Noticing that

$$\sum_{\mu=1}^M \frac{1}{P(\mu|\mathbf{x})} \frac{\partial P(\mu|\mathbf{x})}{\partial x_k} \frac{\partial P(\mu|\mathbf{x})}{\partial x_l} = - \sum_{\mu=1}^M P(\mu|\mathbf{x}) \frac{\partial^2 \ln P(\mu|\mathbf{x})}{\partial x_k \partial x_l} \quad (\text{A.26})$$

we can introduce the  $K \times K$  Fisher information matrix of the categories,  $F_{\text{cat}}(\mathbf{x})$ , characterizing the sensitivity of  $\mu$  with respect to small variations of  $\mathbf{x}$ :

$$[F_{\text{cat}}(\mathbf{x})]_{kl} = - \sum_{\mu=1}^M P(\mu|\mathbf{x}) \frac{\partial^2 \ln P(\mu|\mathbf{x})}{\partial x_k \partial x_l} \quad (\text{A.27})$$

This eventually leads to equation (3.23):

$$I(\mu, \mathbf{x}) - I(\mu, \mathbf{r}) = \frac{1}{2} \int d^K \mathbf{x} p(\mathbf{x}) F_{\text{cat}}(\mathbf{x}) : F_{\text{code}}^{-1}(\mathbf{x}) \quad (\text{A.28})$$

Since  $F_{\text{code}}$  is of order  $N$ , one has in particular that  $I(\mu, \mathbf{x}) - I(\mu, \mathbf{r})$  is of order  $1/N$ .

## B Derivation of Equation (3.28)

Each cell has an activity  $r_i$  equal to 1 if  $x$  is in  $[\theta_i, \theta_{i+1}]$ , and 0 otherwise. The width of the  $i$ th tuning curves is thus  $a_i = \theta_{i+1} - \theta_i$ , and we define the preferred stimuli as the center of the receptive fields,  $x_i \equiv (\theta_i + \theta_{i+1})/2$ .

We want to compute

$$\Delta \equiv I(\mu, x) - I(\mu, \mathbf{r}) = \mathcal{H}(\mu|\mathbf{r}) - \mathcal{H}(\mu|x) \quad (\text{B.1})$$

where

$$\begin{aligned} \mathcal{H}(\mu|x) &= - \int dx p(x) \sum_{\mu=1}^M P(\mu|x) \ln P(\mu|x) \\ \mathcal{H}(\mu|\mathbf{r}) &= - \int d^N \mathbf{r} P(\mathbf{r}) \sum_{\mu=1}^M Q(\mu|\mathbf{r}) \ln Q(\mu|\mathbf{r}) \end{aligned}$$

in terms of the  $\{\theta_i, i = 1, \dots, N+1\}$ , and to see what is the optimal choice of the  $\{\theta_i, i = 1, \dots, N+1\}$ , or equivalently of the  $\{x_i, i = 1, \dots, N\}$ , for having this difference  $\Delta$  as small as possible for a large but finite value of  $N$ .

If we set:

$$\tilde{P}_i \equiv \int_{\theta_i}^{\theta_{i+1}} dx p(x) \quad (\text{B.2})$$

$$\tilde{P}_i(\mu) \equiv \int_{\theta_i}^{\theta_{i+1}} dx p(x|\mu) \quad (\text{B.3})$$

$$\tilde{Q}_{i,\mu} \equiv P(\mu|r_i = 1) = \frac{\tilde{P}_i(\mu) q_\mu}{\tilde{P}_i} \quad (\text{B.4})$$

$$Q_\mu(x) \equiv P(\mu|x) \quad (\text{B.5})$$

then we can write  $\Delta$  from equation (B.1) as :

$$\Delta = \sum_i \Delta_i \quad (\text{B.6})$$

with

$$\Delta_i = \int_{\theta_i}^{\theta_{i+1}} dx p(x) \left[ \mathcal{H}(\{\tilde{Q}_{i,\mu}\}) - \mathcal{H}(\{Q_\mu(x)\}) \right] \quad (\text{B.7})$$

where  $\mathcal{H}(\{Q_\mu\}_{\mu=1}^M)$  is the mixing entropy:

$$\mathcal{H}(\{Q_\mu\}_{\mu=1}^M) = - \sum_{\mu=1}^M Q_\mu \ln Q_\mu \quad (\text{B.8})$$

The quantity  $\mathcal{H}(\{Q_\mu(x)\})$  is zero on a homogeneous domain, hence there is no contribution from intervals included in such domain. Suppose on the contrary that on the full range under consideration  $\mathcal{H}(\{Q_\mu(x)\})$  is rapidly varying. We expect then the optimal  $\theta_i$ 's distribution to be dense, that is  $a_i = \theta_{i+1} - \theta_i$  small.

Recalling that  $x_i = (\theta_{i+1} + \theta_i)/2$  and assuming  $a_i \ll 1$ ,

$$\begin{aligned} \tilde{P}_i &= \int_{-a_i/2}^{a_i/2} dz [p(x_i) + zp'(x_i) + \frac{z^2}{2}p''(x_i)] \\ &= a_i p(x_i) + \frac{a_i^3}{24} p''(x_i). \end{aligned}$$

Likewise,  $\tilde{P}_i(\mu) = a_i p(x_i|\mu) + a_i^3/24 p''(x_i|\mu)$ . Thus,

$$\tilde{Q}_{i,\mu} = Q_\mu(x_i) + \frac{a_i^2}{24} A_{i,\mu} \quad (\text{B.9})$$

with

$$A_{i,\mu} = Q_\mu''(x_i) + 2 \frac{p'(x_i)}{p(x_i)} Q_\mu'(x_i) \quad (\text{B.10})$$

A Taylor expansion then gives :

$$\begin{aligned}\mathcal{H}(\{\tilde{Q}_{i,\mu}\}) &= \mathcal{H}(\{Q_\mu(x_i)\}) + \nabla\mathcal{H}(\{Q_\mu(x_i)\}) \cdot \left( \dots \frac{a_i^2 A_{i,\mu}}{24} \dots \right)^\top \\ &+ \frac{1}{2} \left( \dots \frac{a_i^2 A_{i,\mu}}{24} \dots \right) \mathbb{H}_{\mathcal{H}}(\{Q_\mu(x_i)\}) \left( \dots \frac{a_i^2 A_{i,\mu}}{24} \dots \right)^\top\end{aligned}\quad (\text{B.11})$$

where  $\nabla\mathcal{H}(\{Q_\mu(x_i)\})$  and  $\mathbb{H}_{\mathcal{H}}(\{Q_\mu(x_i)\})$  are respectively the gradient and the hessian of  $\mathcal{H}$  evaluated in  $\{Q_\mu(x_i)\}$  (see (A.21) and (A.11)). A second order expansion leads to:

$$\mathcal{H}(\{\tilde{Q}_{i,\mu}\}) = \mathcal{H}(\{Q_\mu(x_i)\}) - \frac{a_i^2}{24} \sum_{\mu=1}^M A_{i,\mu} (\ln Q_\mu(x_i) + 1) \quad (\text{B.12})$$

Thus we have :

$$\begin{aligned}\int_{\theta_i}^{\theta_{i+1}} dx p(x) \mathcal{H}(\{\tilde{Q}_{i,\mu}\}) &= a_i p(x_i) \mathcal{H}(\{Q_\mu(x_i)\}) + \frac{a_i^3}{24} \left( p''(x_i) \mathcal{H}(\{Q_\mu(x_i)\}) \right. \\ &\quad \left. - p(x_i) \sum_{\mu=1}^M A_{i,\mu} (\ln Q_\mu(x_i) + 1) \right)\end{aligned}\quad (\text{B.13})$$

Moreover,

$$\int_{\theta_i}^{\theta_{i+1}} dx p(x) \mathcal{H}(\{Q_\mu(x)\}) = a_i p(x_i) \mathcal{H}(\{Q_\mu(x_i)\}) + \frac{a_i^3}{24} \frac{\partial^2 p(x) \mathcal{H}(\{Q_\mu(x)\})}{\partial x^2} \Big|_{x_i} \quad (\text{B.14})$$

with

$$\begin{aligned}\frac{\partial^2 p(x) \mathcal{H}(\{Q_\mu(x)\})}{\partial x^2} \Big|_{x_i} &= -2p'(x_i) \sum_{\mu=1}^M Q'_\mu(x_i) (\ln Q_\mu(x_i) + 1) \\ &\quad - p(x_i) \left( \sum_{\mu=1}^M \left\{ Q''_\mu(x_i) (\ln Q_\mu(x_i) + 1) + \frac{Q'_\mu(x_i)^2}{Q_\mu(x_i)} \right\} \right) \\ &\quad - p''(x_i) \sum_{\mu=1}^M Q_\mu(x_i) \ln Q_\mu(x_i)\end{aligned}\quad (\text{B.15})$$

Putting equations (B.13), (B.14) and (B.15) together eventually leads to :

$$\Delta_i = \frac{a_i^3}{24} p(x_i) \sum_{\mu=1}^M \frac{P'(\mu|x_i)^2}{P(\mu|x_i)} \quad (\text{B.16})$$

hence

$$\Delta = \sum_i \frac{a_i^3}{24} p(x_i) F_{\text{cat}}(x_i) \quad (\text{B.17})$$

## References

- Abbott, L. and Dayan, P. (1999). The effect of correlated variability on the accuracy of a population code. *Neural Computation*, 11:91–101.
- Abramson, A. and Lisker, L. (1970). Discriminability along the voicing continuum: Cross-language tests. In *Proceedings of the Sixth International Congress of Phonetic Sciences*. Prague: Academia.
- Ashby, F. and Spiering, B. (2004). The neurobiology of category learning. *Behavioral and Cognitive Neuroscience Reviews*, 3(2):101–113.
- Averbeck, B., Latham, P., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7:358–366.
- Blahut, R. E. (1987). *Principles and practice of information theory*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- Bonnasse-Gahot, L. (in preparation).
- Brunel, N. and Nadal, J.-P. (1998). Mutual information, fisher information, and population coding. *Neural Computation*, 10:1731–1757.
- Butts, D. A. and Goldman, M. S. (2006). Tuning curves, neuronal variability, and sensory coding. *PLoS Biology*, 4(4):e92.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.
- Cover, T. and Thomas, J. (2006). *Elements of Information Theory*. Wiley & Sons, New York. Second Edition.
- Dayan, P. and Abbott, L. F. (2001). *Theoretical Neuroscience*. MIT Press.
- Duda, R. and Hart, P. (1973). *Pattern Classification and Scene Analysis*. Wiley, N.Y.
- Fisher, J. and Principe, J. (1998.). A methodology for information theoretic feature extraction. In Stuberud, A., editor, *Proceedings of the IEEE International Joint Conference on Neural Networks*.

- Freedman, D., Riesenhuber, M., Poggio, T., and Miller, E. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291:312–316.
- Freedman, D., Riesenhuber, M., Poggio, T., and Miller, E. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *Journal of Neuroscience*, 15:5235–5246.
- Georgopoulos, A., Schwartz, A., and Kettner, R. (1986). Neuronal population coding of movement direction. *Science*, 233:1416–1419.
- Goldstone, R. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 123(2):178–200.
- Green, D. and Swets, J. (1988). *Signal detection theory and psychophysics, reprint edition*. Los Altos, CA: Peninsula Publishing.
- Guenther, F., Husain, F., Cohen, M., and Shinn-Cunningham, B. (1999). Effects of categorization and discrimination training on auditory perceptual space. *Journal of the Acoustical Society of America*, 106:2900–2912.
- Han, Y., Köver, H., Insanally, M., Semerdjian, J., and Bao, S. (2007). Early experience impairs perceptual discrimination. *Nature Neuroscience*, 20(9):1191–1197.
- Harnad, S., editor (1987). *Categorical Perception: The Groundwork of Cognition*. New York: Cambridge University Press.
- Harnad, S. (2005). Cognition is categorization. In Cohen, H. and Lefebvre, C., editors, *Handbook of Categorization*. Elsevier.
- Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. (1995). Acoustic characteristics of american english vowels. *Journal of the Acoustical Society of America*, 97(5):3099–3111.
- Hintzman, D. (1986). “schema abstraction” in a multiple-trace memory model. *Psychological Review*, 93(4):411–428.
- Humphreys, G. and Forde, E. (2001). Hierarchies, similarity and interactivity in object recognition: “category-specific” neuropsychological deficits. *Behavioral and Brain Sciences*, 24:453–509.

- Hung, C., G.Kreiman, Poggio, T., and DiCarlo, J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310:863–866.
- Jiang, X., Bradley, E., Rini, R., Zeffiro, T., VanMeter, J., and Riesenhuber, M. (2007). Categorization training results in shape- and category-selective human neural plasticity. *Neuron*, 53:891–903.
- Kang, K., Shapley, R., and Sompolinsky, H. (2004). Information tuning of populations of neurons in primary visual cortex. *Journal of Neuroscience*, 24(13):3726–3735.
- Kang, K. and Sompolinsky, H. (2001). Mutual information of population codes and distance measures in probability space. *Physical Review Letters*, 86(21):4958–4961.
- Kiani, R., Esteky, H., Mirpour, K., and Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, 97:4296–4309.
- Knoblich, U., Freedman, D., and Riesenhuber, M. (2002). Categorization in it and pfc: Model and experiments. *AI Memo 2002-007. Cambridge, MA: MIT AI Laboratory*.
- Kobatake, E., Wang, G., and Tanaka, K. (1998). Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *Journal of Neurophysiology*, 80:324–330.
- Koida, K. and Komatsu, H. (2007). Effects of task demands on the responses of color-selective neurons in the inferior temporal cortex. *Nature Neuroscience*, 10(1):108–116.
- Kruschke, J. (1992). *Alcove* : An exemplar-based connectionist model of category learning. *Psychological Review*, 99(1):22–44.
- Kuhl, P. (1991). Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50(2):93–107.

- Kuhl, P. and Padden, D. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *Journal of the Acoustical Society of America*, 73(3):1003–1010.
- Li, W., Piech, V., and Gilbert, C. (2004). Perceptual learning and top-down influences in primary visual cortex. *Nature Neuroscience*, 7(6):651–658.
- Liberman, A., Harris, K., Hoffman, H., and Griffith, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54:358–369.
- Livingston, K., Andrews, J., and Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 24(3):732–753.
- Logothetis, N., J.Pauls, and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5(5):552–563.
- Minsky, M. and Papert, S. (1969). *Perceptrons*. MIT Press, Cambridge.
- Nadal, J.-P. (1994). Formal neural networks: from supervised to unsupervised learning. In Goles, E. and Martinez, S., editors, *Cellular Automata, dynamical systems and neural networks*, pages 147–166. Kluwer, Book series 'Mathematics and its applications' vol. 282.
- Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology*, 115(1):39–57.
- Op de Beeck, H., Wagemans, J., and Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience*, 4(12):1244–1252.
- Palmeri, T. and Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience*, 5:291–304.
- Paradiso, M. (1988). A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. *Biological Cybernetics*, 58:35–49.
- Poggio, T. (1990). A theory of how the brain might work. *Cold Spring Harbor Symp. Quant. Biol.*, 55:899–910.

- Poggio, T. and Girosi, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497.
- Pouget, A., Zhang, K., Deneve, S., and Latham, P. (1998). Statistically efficient estimation using population coding. *Neural Computation*, 10:373–401.
- Rieke, F., Warland, D., de Ruyter van Steveninck, R., and Bialek, W. (1997). *Spikes: Exploring the Neural Code*. MIT Press, Cambridge.
- Riesenhuber, M. and Poggio, T. (2000). Models of object recognition. *Nature Neuroscience*, 3:1199–1204.
- Schölkopf, B., Burges, C., and Smola, A., editors (1999). *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge.
- Seriès, P., Latham, P., and Pouget, A. (2004). Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nature Neuroscience*, 7(10):1129–1135.
- Seung, H. S. and Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proceedings of the National Academy of Science*, 90:10749–10753.
- Sigala, N. (2004). Visual categorization and the inferior temporal cortex. *Behavioural Brain Research*, 149:1–7.
- Sigala, N. and Logothetis, N. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, 415:318–320.
- Softky, W. and Koch, C. (1993). The highly irregular firing of cortical cells is inconsistent with temporal integration of random epsps. *The Journal of Neuroscience*, 12(1):334–350.
- Sompolinsky, H., Yoon, H., Kang, K., and Shamir, M. (2001). Population coding in neuronal systems with correlated noise. *Physical Review E*, 64(5):051904.
- Stein, R. (1967). The information capacity of nerve cells using a frequency code. *Biophysical Journal*, 7:797–826.
- Sugase, Y., Yamane, S., Ueno, S., and Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, 400:869–873.



- Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19:109–139.
- Taube, J., Muller, R., and J.B. Ranck, J. (1990). Head-direction cells recorded from the postsuiculum in freely moving rats. i. description and quantitative analysis. *The Journal of Neuroscience*, 10(2):420–435.
- Thomas, E., Hulle, M. V., and Vogels, R. (2001). Encoding of categories by noncategory-specific neurons in the inferior temporal cortex. *Journal of Cognitive Neuroscience*, 13(2):190–200.
- Tolhurst, D., Movshon, J., and Dean, A. (1983). The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Research*, 23:775–785.
- Torkkola, K. and Campbell, W. M. (2000). Mutual information in learning feature transformations. In *Proc. 17th International Conf. on Machine Learning*, pages 1015–1022. Morgan Kaufmann, San Francisco, CA.
- Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. part 2: single-cells study. *European Journal of Neuroscience*, 11:1239–1255.
- Vogels, R. and Orban, G. (1990). How well do response changes of striate neurons signal differences in orientation: A study in the discriminating monkey. *The Journal of Neuroscience*, 10(11):3543–3558.
- Wilson, M. and DeBauche, B. (1981). Inferotemporal cortex and categorical perception of visual stimuli by monkeys. *Neuropsychologia*, 19(1):29–41.
- Yoon, H. and Sompolinsky, H. (1999). The effect of correlations on the fisher information of population codes. In Kearns, M., Solla, S., and Cohn, D., editors, *Advances in neural information processing systems 11 (NIPS-11)*, pages 167–173. The MIT Press.
- Young, M. and Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science*, 256:1327–1330.