Local metrical information: application to the perceptual magnet effect

Romain Brasselet and Angelo Arleo
CNRS-UPMC Univ. Paris 6, UMR 7102
F75005, Paris, France
romain.brasselet@upmc.fr

ABSTRACT
In this paper, we propose an extension of the recently introduced metrical information [1]. We give a local version of it, where the similarity between two events can depend on the position in the event space. While the original method demands the equivocation (or conditional entropy) to be zero and then maximize the information, we here define an objective function requiring a trade-off parameter between the minimization of equivocation and the maximization of information. We show that the well-known perceptual magnet [2] effect can be understood as a consequence of maximizing this objective function in a low-noise regime. In a simple case with narrow gaussian categories, it implies that an optimal decoding system should perform very fine discrimination near the boundaries of the categories while being less demanding near the centers of the categories. This mechanism could be implemented as exhibited in [3]. In the spirit of [4], we here show how the optimal decoder should behave in the low and high-noise regimes.

KEY WORDS
Information theory, neurotransmission, temporal neural code, spike-train metrics

1. Introduction

Shannon information theory is one of the best framework to estimate the amount of information that is conveyed by a noisy transmission system. It tells how much a decoder can possibly know about the source by reading the message transmitted. In the field of neuroscience, it has been extensively used [5, 6, 7]. However, readout neurons have properties that make them depart from the ideal observer, properties that may be formalized by metrics. In addition, there may be details of the code that it is not relevant to convey. A recent proposal [1] -akin to the bottleneck information [8] and to Relevant Component Analysis [9]- was to merge the metrics with information theory so as to find the metrics (and thereby the properties of the neurons) that make them transmit the most relevant information. We propose to generalize the previously introduced metrical information by allowing the similarity function to depend on the position in the event space.
An information-theoretic measure embedding the metrical relations between the events was recently introduced to link the information conveyed by a neurotransmission system on account of the properties of the downstream decoder. This measure is based on an entropy that can be written:

$$H*(X) = -\sum_x p(x) \log \sum_y p(y)\varphi(x,y)$$

where $\phi(x,y)$ is a similarity measure between the events x and y that depends on the distance between them. The distance and the similarity measure are thought of as representations of the properties of the downstream decoder. In the original definition, the similarity measure was taken as an all-or-none function of the distance with a cut-off value called the critical distance. It was then proposed that an optimal decoder should at the same time minimize the equivocation (conditional entropy) and maximize the information so as to be able to reconstruct the stimulus unambiguously. This could be done by choosing an appropriate definition of the distance and a well-tuned critical distance.
We wish to make a remark concerning this definition. It is global: the similarity measure does not depend on the location on the event space. This means that the discrimination capacity is the same over all the space. We here introduce a local version of this metrical entropy. We then show that the attempt to maximize the information and minimize the equivocation can be formalized as maximizing a single objective function that takes a trade-off parameter as input. When applying this optimization to gaussian distributions, we found results very much similar to the physiological phenomenon of perceptual magnet.

2. Methods

The metrical information is a reformulation of Shannon Mutual Information that takes into account the similarity between objects. It may come in two versions: one is global in the sense that the similarity measure is invariant under translations in the output space. In the local version, the similarity measure may depend on the output. This is the latter version that will be used henceforth. In this version, the information (resp. equivocation) is maximum (resp. minimum) if the outputs of a category are all closer than they are from the outputs of the other categories. Indeed, in that case, a similarity measure at a given point includes all the responses to the same category while excluding all other responses. Note that here, the similarity measure can be thought of as the processing of the inputs by the nervous system. It is symbolic in the sense that we do not claim it should be implemented at some precise location, but could be the result of a multi-stage

processing. The crucial variables in these measures are the similarity measures at each point in the response space. In case one wants to optimize the discrimination of categories, i.e. maximize the hit rate and minimize the false alarm rate, how should these variables be chosen? Here, instead of only trying to maximize this information, we will look for conditions that maximize an objective function defined as:

$$Q(R,S) = I*(R,S) - \alpha H*(R|S) \qquad (1)$$

If α = 0, the objective function is just the information. The factor α determines your main aim: minimizing the metrical equivocation or maximizing the metrical information. Thus, we do not want an observer that acts a the perfect Shannon observer by differentiating as much as he can by taking into account every slight differences between responses. The problem is more that of finding the characteristics of an observer who would really reconstruct the categories.

Since we now consider a local version (where the similarity function depends on the position on the space), we define the specific information given by a response r:

$$i*(r,S) = \sum_s p(r|s) \log \frac{\sum_{r'} p(r'|s)\varphi(r,r')}{\sum_{r'} p(r')\varphi(r,r')} \qquad (2)$$

and the specific equivocation:

$$h*(r|S) = -\sum_r p(r|s) \log \sum_{r'} p(r'|s)\varphi(r,r') \qquad (3)$$

and thus a specific objective function:

$$q(r,S) = i*(r,S) - \alpha h*(r|S)$$

3. Results

Let us consider a simple case with a few categories and responses lying in the set of real numbers. For simplicity, let us assume a Gaussian distribution for each object i, with mean μi and standard deviation σi . Let us simplify even more by considering the variances to be equal, σi = σ, ∀i. Henceforth, optimal discrimination is studied by considering the objective function defined in Eq. 1 and by varying the trade-off parameter α. Recall that a large α implies maximizing the metrical information and minimizing the metrical equivocation simultaneously, whereas small α values relax the minimality constraint on the metrical equivocation. We will consider Gaussian similarity kernel functions ϕ(x,y) = exp((x−y)^2/2β^2) , which are more realistic than Heaviside-like all-or-none functions. Different standard deviations β will be employed in order to modulate the selectivity of the similarity function (the lower β, the more selective ϕ). In particular, we will allow β to vary as a function of space to see what regions of the input space should be discriminated more than the others. Figure 1 presents the results obtained with four different values of the trade-off parameter, α =0, 10^{-7} , 10^{-5} , 1, in the presence of low-noise category distributions (blue curves). The optimal width β of the local similarity kernel is shown at each value of the response space (black curves). The

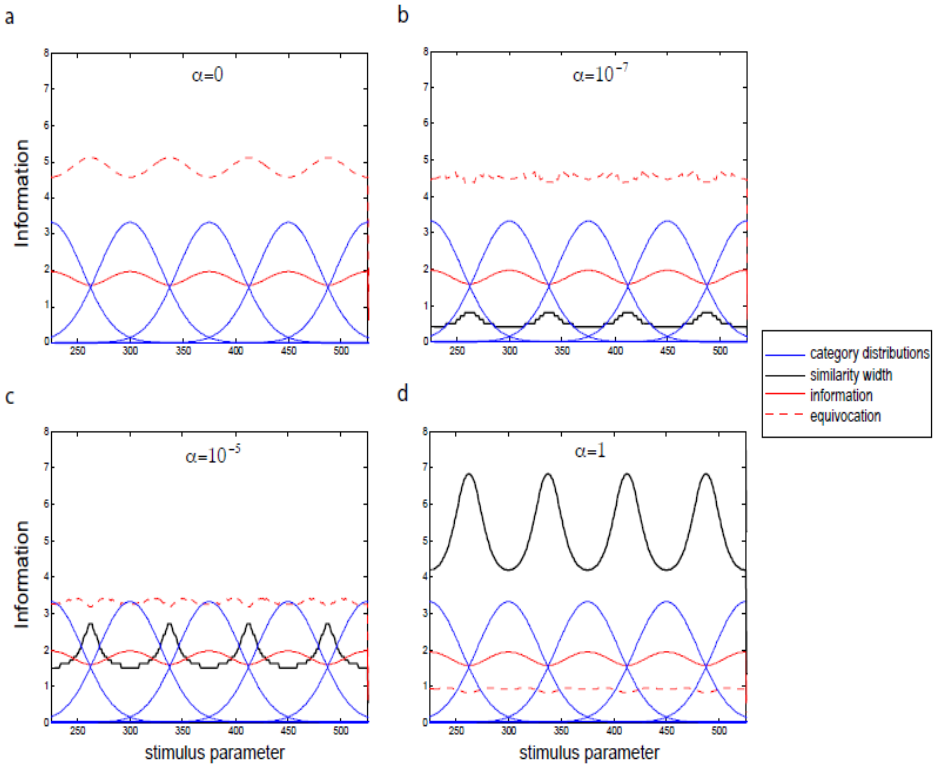

Illustration 1: Optimal discrimination in high-noise regime. Five categories (blue) are to be discriminated. The results were obtained for 4 different values of the trade-off parameter α =0, 10^{-7} , 10^{-5} , 10^{-1} . For all values of α, the width is largest in-between categories.

specific information $i*(r;S)$, Eq. 2, and equivocation $h*(r|S)$, Eq. 3, are also plotted (solid and dashed red curves, respectively) to show the contribution of all the responses to the expectation values over the entire space. For $\alpha = 0$ (Fig. 1a), the goal is to maximize the information without taking into account the equivocation. The best solution is to have very narrow similarity functions so as to discriminate every pairs of responses: the width β is thus always zero. For low (positive) values of α, the width of the similarity function can be large at the center of a category distribution since it decreases the equivocation without impinging on the information. This is apparent in Fig. 1b, corresponding to $\alpha = 10^{-7}$, where the values of the equivocation drop at the center of the categories compared to the $\alpha = 0$ case. On the other hand, to keep the information maximum, the kernel function ϕ must be narrow in-between categories (increasing the selectivity of the decoder). Indeed, a higher selectivity in-between categories guarantees a high information, at the price of a large equivocation, since objects at the center of a category will not be considered similar to those at the edge. For $\alpha = 10^{-5}$, Fig. 1c, the equivocation is already almost nil at the center of the categories (with a full information), though it is far from being zero in-between them. However, the contributions to the equivocation of these events are small because their probability is low. The equivocation further decreases only when α is drastically increased (Fig. 1d, $\alpha = 1$). The price to pay is a loss of information in-between categories since now, objects belonging to some category are considered similar to objects of other categories.

We here see the effect of the trade-off parameter at each point: in general, a response near the edge of a category contributes largely to the equivocation as long as its similarity with the center of the category (which has a high probability) is low.

However, increasing its width impinge on the information. At high values of the trade-off parameter (e.g. $\alpha = 1$), the similarity measure is wider in-between categories than at their centers. The reason is that, for the specific equivocation to be zero (i.e. all objects from the same category considered identical), the similarity measure at the boundaries of a category needs to be twice as wide as that at the center of a category. If we arbitrarily give Shannon information and equivocation values of 1, the metrical information and equivocation are respectively:

α	$I*(R,S)$	$H*(R\|S)$
0	1	1
10^{-7}	1	0.63
10^{-5}	1	0.41
1	0.96	0.05

which shows that, in this case, it is feasible to decrease the equivocation without impinging on the information. Figure 2 shows that in the high-noise regime (large variance of the distributions) the optimal cut-off distances behave differently than in the low-noise regime. The optimal widths β of the similarity kernels tend to be large at the interface between classes and rather small near the centers. The situation in which the similarity measures are wide at the center of categories never occurs in the high-noise regime. As a onsequence, it is almost impossible to decrease the equivocation without diminishing the information.

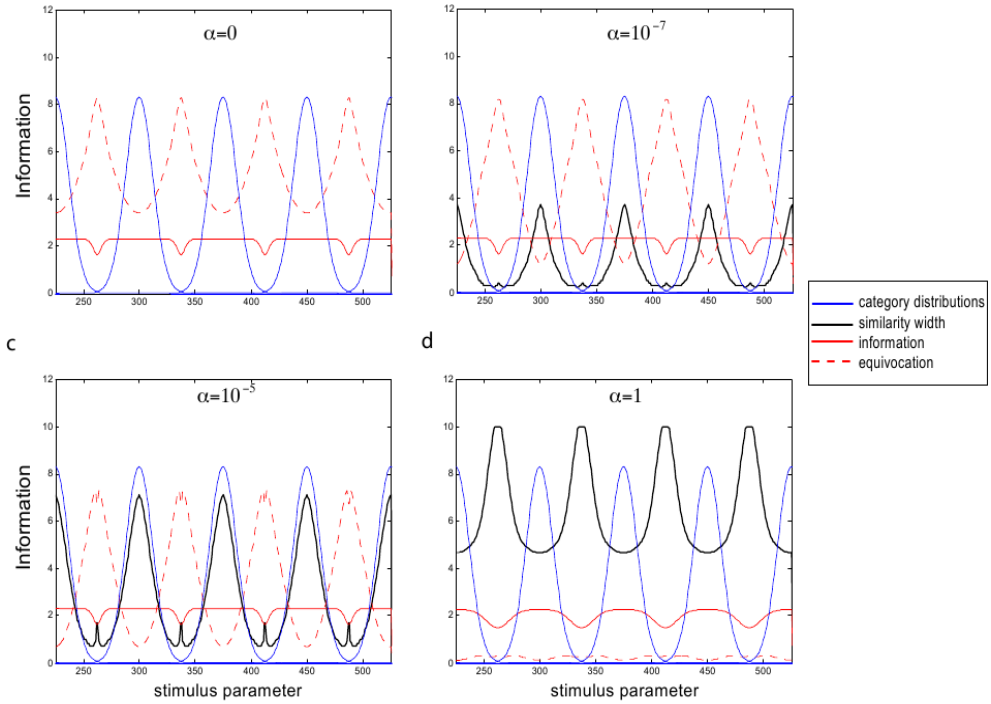

Illustration 2: Optimal discrimination in low-noise regime. Five categories (blue lines) must be discriminated. The results were obtained for values of the trade-off parameter $\alpha = 0, 10^{-7}, 10^{-5}, 1$. For low values of α, the width β of the similarity measure at each point (black curve) is larger at the center of the categories and smaller between them (indicating lower and higher selectivity, respectively). As the trade-off parameter α increases, the width becomes highest in-between categories. At $\alpha = 10^{-5}$, the information (resp. equivocation) is still high (resp. already low) at the center of the categories.

If the Shannon information and equivocation are attributed values of 1, the information and equivocation are respectively:

α	I ∗ (R, S)	H ∗ (R|S)
0	1	1
10^{-7}	1	0.94
10^{-5}	1	0.69
10^{-1}	0.98	0.19

The results of Figs. 1,2 suggest that for low values of both noise level and trade-off parameter α, a phenomenon of high selectivity in-between categories and low selectivity at their centers occurs. For other values of these two parameters, optimal selectivity is found to be higher at the center of categories.

4. Discussion

4.1 The perceptual magnet

When attempting to maximize the objective function defined with respect to the metrical entropy on a set of Gaussian categories, two types of behavior were observed: the selectivity of the similarity measure can be larger either at the center of the distribution of each category or between categories.

The first scenario can be linked to the perceptual magnet effect [2], which has been first introduced in the field of psychoacoustics in the early 90s. Each category of sound (for example each type of vowel) has a prototype, an element that best represents the vowel. The other elements of the category are perceived closer to the prototype than their physical distance would suggest, as if the prototype was pulling the elements of the category toward itself, hence the term 'perceptual magnet'. The perceptual space is thus warped with between-category expansion and within-category compression.

The results presented here in the low-noise regime for low values of the trade-off parameter α can then be related to the perceptual magnet. Indeed, the width of the similarity measure at the center of a category is large, meaning that objects around are seen as very similar, while the width is small in-between categories, making the close objects dissimilar. It is interesting to highlight the fact that the perceptual magnet may not be a highly generic phenomenon, but rather appear in limited ranges of parameters only. The above results suggest that it seems to be a low-noise regime effect in circumstances where the emphasis is on discrimination of different stimuli rather than identification of similar ones. We studied in more details the predominance of the two types of behavior (i.e. highest width inside categories or in-between them). We defined the amplitude of the perceptual magnet as the logarithm of the ratio between the width of the similarity measure at the exact middle between categories and at the exact center of categories. A highly negative value of this amplitude would correspond to a very pronounced perceptual magnet effect. The phase diagram of Figure 3 displays the amplitude of the perceptual magnet effect with respect to the noise level and to the trade-off parameter α.

The varying width of the similarity measure can be understood in two different ways. The first one is that the input space is isometrically represented in the nervous patterns and some neural decoder can perform the task of the similarity measure. Yet, it could also be the result of a remapping through an ensemble of neurons with a non-uniform density followed by a decoder using a uniform similarity measure. Bonnasse-Gahot and Nadal (2008) [3] presented a study based on Fisher information analysis and showed that a way to implement this mechanisms may be to have a higher receptive field density between the categories than within categories. In the light of metrical information analysis, this possible neural coding principle can interpreted the following way. The input space is mapped to the space of discharge rate of the neurons (i.e. tuning curves). In this space, categories are much more separated with large no man's land between them. Then using a uniform similarity measure on this space yields the maximum objective function. Here, the optimal similarity measure is obtained in two steps: first an optimal neural coding that relatively contracts the categories with respect to the distances between them and then a decoding scheme (whose implementation is not given) that maximizes the quality factor. The simple example of Figure 3 suggest that the two approaches (Fisher and metrical information) yield similar results.

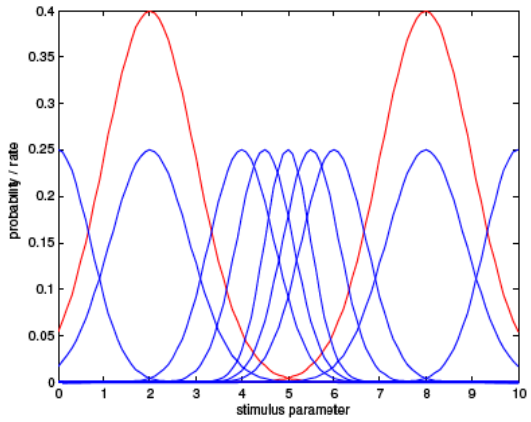

Illustration 3: Example of 2 categories (red curves) and a population of tuning curves (blue lines). In the response space of the population of neurons, the distance between x = 2 and x = 4 is equal to d1 ∼ 2.8, while that between x = 4 and x = 6 is d2 ∼ 3.8. Therefore, the categories are much smaller compared to their mutual distances.

4.2 Phase diagram

It is interesting here to highlight the fact that the perceptual magnet may not be a highly generic phenomenon, but rather appears in limited ranges of

parameters. In particular, it seems to be a low-noise regime effect in circumstances where the emphasis is on discrimination of different stimuli (high information) rather than identification of similar ones (low equivocation).

We studied in more details the predominance of the two types of behaviour: highest width inside categories or in-between them. We defined the amplitude of the perceptual magnet as the logarithm of the ratio between the width of the similarity measure at the exact middle between categories and at the exact center of categories. A highly negative value of this amplitude corresponds to a very pronounced perceptual magnet effect. The amplitude is plotted with respect to the noise level and to the trade-off parameter in fig. 4.

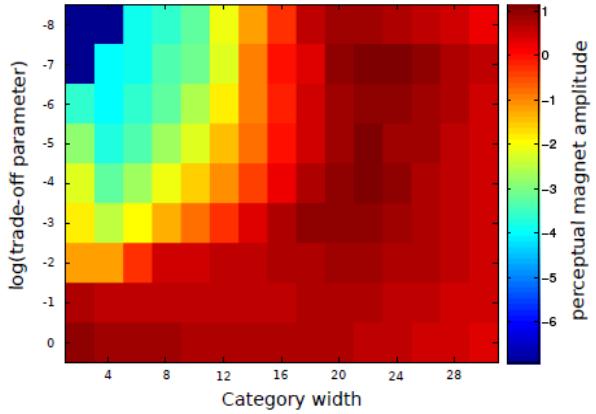

Illustration 4: Perceptual magnet phase diagram. The perceptual magnet amplitude is defined as the logarithm of the ratio of the width of the similarity measure between categories and at the center of categories. This amplitude is plotted with respect to the noise (category width) and to the trade-off parameter. The perceptual magnet appears to occupy only a small portion of the phase space: in the low-noise regime and small trade-off parameter.

5. Conclusion

We have shown that the metrical information yields results that can be related to the perceptual magnet effect. It theoretically allows us to give estimates on the presence of this phenomenon in a two-dimensional phase space consisting of the noise level and the goal of the communication system (maximizing true positives or minimizing false positives). In the low noise regime, approaches based on Fisher information [3] yielded similar results if we make a straightforward link between the firing rates of a population of neurons and the distances they implement. It is indeed known that Fisher information is somewhat dependent on the metrics of the event space. This observation explains in part why both approaches yield similar results. But, more than this, we think that future work may help clarifying further the relations between metrical information and Fisher information and thus bring light on the relations between classical information theory and Fisher information [10].

Acknowledgements

The authors thank Laurent Bonnasse-Gahot and Jean-Pierre Nadal for stimulating discussion. R.B. thanks the Fondation pour la Recherche Médicale.

References

[1] R. Brasselet, R.S. Johansson, and A. Arleo. Optimal context separation of spiking haptic signals by second-order somatosensory neurons. *In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, Advances in Neural Information Processing Systems 22, pages 180–188. 2009.*

[2] P. Kuhl. Human adults and human infants show a perceptual magnet effect for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics, 50(2):93–107, 1991.*

[3] L. Bonnasse-Gahot and J. P. Nadal. Neural coding of categories: information efficiency and optimal population codes. *Journal of Computational Neuroscience, 25(1):169–187, 2008.*

[4] D.A. Butts and M.S. Goldman. Tuning curves, neuronal variability, and sensory coding. *PloS Biol, 4:e92, 2006.*

[5] W. Bialek, F. Rieke, R. R. de Ruyter van Steveninck, and D. Warland. Reading a neural code. *Science, 252:1854–1857, 1991.*

[6] A. Borst and F. E. Theunissen. Information theory and neural coding. *Nat. Neurosci., 2:947– 957, 1999.*

[7] R. Quian Quiroga and S. Panzeri. Extracting information from neuronal populations: information theory and decoding approaches. *Nat. Rev. Neurosci., 10:173–185, 2009.*

[8] N. Tishby, W. Bialek, and F.C. Pereira. The information bottleneck method. *In Proceedings of the 37th annual Allerton conference on communication, control and computing, 1999.*

[9] N. Shental, T. Hertz, D. Weinshall, and M. Pavel. Adjustment Learning and Relevant Component Analysis, *pages 776–790. Computer Vision -ECCV.*

[10] N. Brunel and J.-P. Nadal. Mutual information, fisher information, and population coding. *Neural Computation, 10:1731–1757, 1998.*