Neuronal Goals: Efficient Coding and Coincidence Detection

Nathan Intrator*
School of Mathematical Sciences
Tel Aviv University
nin@cns.brown.edu

Abstract— **Barlow's seminal work on minimal entropy codes and unsupervised learning is reiterated. In particular, the need to transmit the probability of events is put in a practical neuronal framework for detecting suspicious events. A variant of the BCM learning rule [15] is presented together with some mathematical results suggesting optimal minimal entropy coding.**

Key words: Sparse coding, Non-Gaussian distributions, BCM Theory, Minimal Entropy

1 Introduction

There is no doubt that much of what we do is determined by what has happened in the past. In particular, our ability to understand speech in noisy environment, understand under contextual constraints, or drive a car, is a manifestation of our ability to predict the next phoneme/word in a sentence, predict the next required control movement, or at least adjust our expectations according to the past context. Clearly, this is a fundamental concept without which the system would not function at all, or would be severely degraded.

It is largely assumed that if the role of sensory neurons is to detect features in their input representation, then they should transmit the probability of occurrence of the features they learn to detect. While this sounds very natural and simple, we argue that such coding is not optimal and in fact neurons can and should transmit additional information.

Following Barlow's seminal work on minimal entropy codes and unsupervised learning, we attempt to address some fundamental problems concerning neuronal coding, neuronal goals for learning, feature detection and information transmission. In particular, the need to transmit the probability of events is put in a practical neuronal framework for detecting suspicious events. We derive these assertions from basic principles of information theory, from energy conservation considerations, and from some assumptions about neuronal goals.

Several other researchers have been interested in these questions. Atick [1] studied information coding patterns in flys and mammalians retinal coding and supports the notion of redundancy reduction through effective information coding. Field et al. [10, 21] inferred about the goal of visual sensory coding from properties of the statistics of natural images. Their main conclusions are the need to extract higher order statistics (i.e., more than linear and pairwise) and the need for sparse coding as a mean to achieve efficient information relay.

In this paper we present a unifying theory that combines the need for efficient feature detection with the need for efficient information transmission and a fundamental neuronal goal for suspicious coincidence detection. We start with a review of Barlow's work on coincidence detection, continue with a review of a BCM neuron in terms of its feature detection ability and information coding properties, and later discuss some regularization properties of these neurons in terms of the probability of events they can become selective to and outlier avoidance. We conclude with a motivation from the principle of maximum entropy to the optimality of the code.

2 Neuronal goal: Suspicious coincidences detection

Barlow has been arguing for a long time that *suspicious coincidences* is the basic type of event to which the cerebral cortex must attune itself [2, 3, 6, 7]. Assuming that a major task of the brain is to form a statistical model of the world, Barlow asked what kind of events would be worth noting and keeping a record of. Clearly, neither isolated events (the falling of a stone) nor repeated occurrences of events (the ticking of a clock) deserve paying too much attention to. In contrast, a co-occurrence of two events may call for investigation or may justify remembering, but only if this co-occurrence is *surprising* (i.e., unlikely), given prior knowledge regarding the occurrence of the individual events. Coincidence detection is also a key idea in the *Compositional Machine* framework presented by Geman and Bienenstock [13].

Consider the statistical problem of learning which tries to determine whether a compound event such as C followed by U is a random co-occurrence or a significant association. If it is the latter then C is a conditional stimulus to U, can be useful in predicting it, and in some cases can be useful in detecting the event U out of several concurrently occurring events. Clearly, we can not determine anything about the combination of C and U before becoming independently selective to each of the events, and having

*Current address: Institute for Brain and Neural Systems, Box 1843, Brown University, Providence, RI 02912

estimated their prior probabilities. Hence, Barlow hypothesises that the perception of an event, not only should signal its occurrence, but must also indicate the prior probability of what has been signaled.

2.1 Selfridge's Pandemonium and Barlow's Probabilistic Pandemonium

The probabilistic line of reasoning suggests that sensory coding is "... the process of preparing a representation of the current sensory scene in a form that enables subsequent learning mechanisms to be versatile and reliable" [6]. Specifically, a representation is useful for learning if it includes records of recurring and co-occurring events. As noted by Barlow, a convenient substrate for such a representation is provided by Selfridge's Pandemonium [23]. In Barlow's Probabilistic Pandemonium, the response strength of a feature-detector demon would be proportional to $-\log P$, where P is the probability of occurrence of the feature the demon detects. These signals are then propagated to an association network which receives *unconditional* inputs as well – inputs that follow and are assumed to be related to the *conditional* input. The main innovation in this setup, is the argument that each demon (feature detector) propagates information inversely proportional to the likelihood of the feature that is detected. This is sharply different than more conventional feature detectors, such as say Principal Components, in which the output is proportional to the degree of similarity between the input and the feature (when extracting PC from the correlation of inputs matrix), or the amount of variance explained by that feature (when extracting PC from the covariance matrix).

In the next section we present the feature extraction properties of a BCM neuron and emphasize its coding properties and relevance to coincidence detection.

3 The BCM feature extraction and coding

The feature extraction method briefly described below achieves dimensionality reduction by seeking features that would best distinguish among the members of the set. The potential importance of these features is related to their invariance properties, or their ability to generalize. Invariance properties of features extracted by this method have been demonstrated previously in various recognition tasks [14, 16, 17].

From a mathematical viewpoint, extracting features from gray level images is related to dimensionality reduction in a high dimensional vector space, in which an $n \times k$ pixel image is considered to be a vector of length $n \times k$. The dimensionality reduction is achieved by replacing each image (or its high dimensional equivalent vector) by a low dimensional vector in which each element represents a projection of the image onto a vector of synaptic weights.

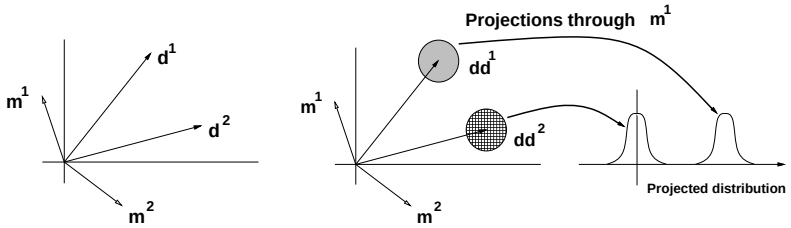

Figure 1: The stable solutions for a two dimensional two input problem are m_1 and m_2 (left) and similarly with a two-cluster data (right).

The BCM feature extraction [8, 15] seeks multi-modality in the projected distribution of these high dimensional vectors. A simple example is illustrated in Figure 1. For a two-input problem in two dimensions, the stable solutions (projection directions) are m_1 and m_2, each has the property of being orthogonal to one of the inputs. In a higher dimensional space, for n linearly independent inputs, a stable solution is one that is orthogonal to all but one of the inputs. In case of noisy but clustered inputs, a stable solution will be orthogonal to all but one of the cluster centers. As is seen in Figure 1 (right), this leads to a bimodal, or, in general, multi-modal, projected distribution.

The mathematical results concerning the type of feature detection and coding is given in [15]. One of the results roughly says:

Theorem With n clusters in an n-dimensional space, the only stable solutions are projections which are orthogonal to all but one of the clusters. There are at most n such solutions and each such solution occurs with probability P_i (the probability of cluster i). Moreover, the neuronal activity of a neuron that becomes tuned to cluster i (in the linear case) is $1/P_i$.

This result makes the BCM neuron a good candidate for efficiently coding events. By adding a monotone truncated log function on top of the neuronal activity, we get the desired -log probability of events (see the the discussion of the optimality of this code below). The truncation at zero is required as neuronal activity is expected to be non-negative.

While the activity of a neuron becomes close to the inverse of the probability of the event this is not

always the case. If the event is only partially detected, namely, the probability that the event appears in the input at a certain time is not close to one, then the activity of the neuron is degraded accordingly.

3.1 Interplay between suspicious coincidence and avoiding outliers

The above results points at a potential weakness of the BCM neuron: sensitivity to outliers. Clearly, if a neuron becomes tuned to an event with vanishing probability (although the probability of such case is going to zero as well) its activity may grow unbounded. This fact had motivated us in the past to apply a saturating sigmoidal transfer function to the neuronal activity [15]. Such saturation function should have an upper bound that is larger than 1, and in fact the upper bound will determine the smallest probability of events that the neuron can become tuned to. The ability to control the probability of events the neuron can become tuned to is very important; It is likely that when detailed low-level features are required, the neuron should be able to detect events with very low probability, but when a high degree of generalization is needed, the neuron should not become tuned to events with very low probability.

Once avoidance of outliers is assured, we can add the monotone log function on top of the neuronal activity for efficient information relay.

4 A Coincidence detection network

In this section, we present a simple architecture that can serve as a coincidence detection network (CDN) and is based on the BCM neurons described above. The first layer of neurons (Figure 2) is composed of feature detectors of events A_i in the input representation. Without loss of generality, we assume that when the dynamic process of learning has stabilized, neuron i becomes selective to event A_i such that the maximal activity of the neuron is around $-\log(P_i)$, where P_i is the probability of event A_i. The second

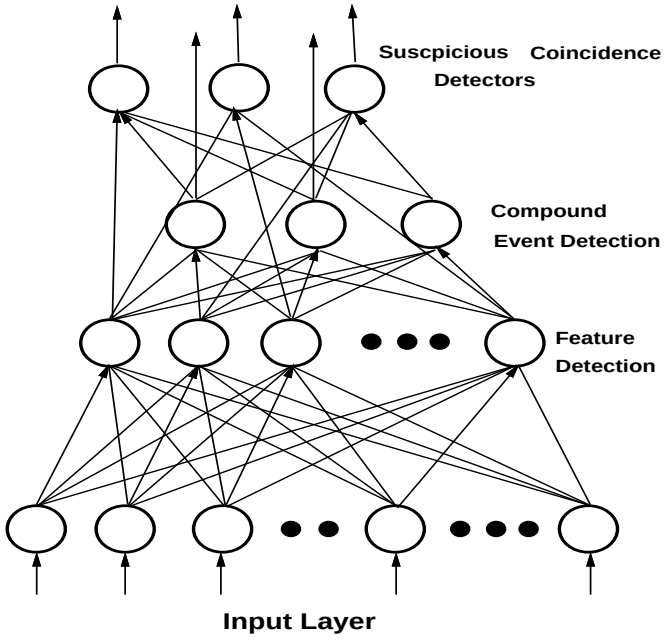

Figure 2: A suspicious coincidence detection network.

layer of neurons receives input from the previous layer, or from other sensory maps, and thus it can only detect events which have been found interesting by earlier feature detectors. Thus, the inputs to the second layer are much more quiet than the inputs to the previous layer, as the events are now sparsely coded, and compound events are more likely to be found. The second layer of BCM neurons again looks for multi-modal projected distribution, which indicate clusters in the activity of the first layer.

With additional sensory input (possibly from other modalities), events A_i and A_j may become correlated and can be detected by the second layer of neurons (e.g. red apple), then emerging projections in the second layer will generate cell activity that is of the form

$$-\sum_j w_j \log P(A_j) = -\log[\Pi_j P(A_j)^{w_j}],$$

such that the resulting neuronal activity represents $-\log P(B)$, where B denotes the compound event. The BCM rule for synaptic weight modification effectively seeks such projections along which the probability density deviates maximally from a Gaussian distribution.[1] The receptive fields of units trained

[1] Due to the central limit theorem, most projections are Gaussian, and thus can be described completely by their

with the BCM rule are thus tuned to the detection of interesting low-dimensional structure in the high-dimensional input space.

5 Minimum entropy coding

Barlow considered the question of "what properties should a representation have in order to make it suitable for use by subsequent learning mechanisms". He argues that not every complete representation would do, and in particular, a model based on Hebbian synapses can only access some of the information needed, but not all. There is an apparent contradiction between the desire to have redundant coding which is related to the simplicity of the code and the need for redundancy reduction as a mean for efficient transmission. Barlow stresses that a completely non redundant stimuli is indistinguishable from random noise [5], thus, requiring a highly sophisticated scheme (probably complex and slow) to decode the signal. Since neuronal code seems highly structured, one can infer that its encoding is highly redundant.

Minimal entropy codes satisfy the need to know the prior probability of events described by the code, and thus, if the variables of the sensory representation occur independently of each other, it is then possible to derive the prior probability of any logical function of these variables from prior probabilities of the individual variables. Therefore, such a representation allows a simple search for suspicious coincidence of events. More specifically, suppose we have a set of neurons tuned to bars at all orientations. A priori, one can assume that the different orientations are independent, but for a specific image, say the character 'A', we create a model by noting that several orientations are highly probable at certain locations of that object. Barlow goes on to describing a *minimum entropy coding* which should satisfy two constraints, it has to be reversible, namely reconstruction of the original input should be possible, and in addition the code should have minimal entropy. The essence of minimal entropy coding, is to form a factorial code, i.e., to find a set of symbols such that each of them occurs independently of the others, so that their joint probability is a multiplication of their individual ones. In the BCM network factorial coding of events, namely the independence of features detected by different neurons is achieved by the lateral inhibition architecture and reduced space of solutions of BCM neurons. The reduced space of solutions is due to the fact that for data with n clusters, there are only n stable solutions, each characterized by a synaptic vector being orthogonal to all but one of the clusters. This space of solutions is minimal and sufficient for distinguishing between n clusters. In contrast, a discriminant analysis method for separation between n clusters [22, for review], has $\binom{n}{2}$ possible solutions, i.e., on the order of n^2, and is thus, more likely to find correlated solutions.

Sparse coding, in which different states of the system are represented by neuronal activity with only a small number of active units, is an outcome of the dynamics of the BCM learning. This is because each BCM neuron conveys events with an activity that is inversely proportional to their probability of occurrence [15]. Thus, in accordance with Barlows predictions [4], events that occur with high probability are conveyed by a less active neuron than events which occur less frequently. Such events will be conveyed by a neuron that is quiet most of the time, but fires strongly when the event is detected. It follows that in the BCM coding case, sparse coding is an outcome of the other constraints and not a direct goal by itself. It will be interesting to compare the resulting code with methods that maximize sparsity or kurtosis as a goal for neuronal coding and feature detection [11, 12, 21].

6 Optimal neuronal code

When seeking optimal neuronal code, we have to bear in mind that the code should preserve spatial relations between the inputs. In particular, two events that are close to one another in measurement space, e.g. two views of the same person, should have an internal representation that preserves this correspondence [9, 25]. This requirement nullifies the use of classical coding theory, since we no longer can use a look-up table that translates the code into symbols to be conveyed from layer to layer, as the symbol representation does not preserve the metric of the original space. One can assume that the additional continuity constraint will limit the optimality and information capacity of codes generated by such a map.

We are now in the following situation: There is a measurement space (X, P) of vectors in R^k. We seek a continuous function on a compact domain in R^k,

$$fC^k :\mapsto R^+,$$

that conveys as much information about the measurement space as possible. The function f is found by learning from a set of observations $\{x_1, \ldots, x_N\}$ sampled from X with probability P. We assume that the activity of a cell is a function of its vector of synaptic weights and the inputs. This defines a distribution over the possible values of cell activity. We discretize those values to a given accuracy and thus are assuming that the collection of cell activity values is given by $f_i,\ i = 1, \ldots, n$, with corresponding probabilities $p_i,\ i = 1, \ldots, n$. The mean cell activity is given by

$$\bar{f} = \sum_{i=1}^{n} p_i f_i, \quad \sum_i p_i = 1, \tag{1}$$

covariance matrix (second-order statistics).

We ask the following question: what is the distribution that maximizes the information capacity of neurons subject to the constraint of a fixed average activity. We assume that the energy dissipated by the neuron is linearly related to the neuronal activity, and thus would like to study the information capacity of neuronal codes with a fixed average activity. We additionally assume that cell activity is non negative, so an inactive cell which dissipates the least amount of energy has zero activity. It turns out that the *principle of maximum entropy* [18, 19, For discussion], is directly applicable in this case. It gives an explicit relation between neuronal activity (possibly firing rate) and the corresponding probability for this activity level, so as to maximize information capacity subject to a fixed mean activity.

At a first glance, it is not intuitive at all that there is any such connection. We often study the information entropy of distribution which is given by

$$H(p_1, \ldots, p_n) = -K \sum_i p_i \ln p_i, \tag{2}$$

where K is a positive constant. However, this case is only applicable when the actual events to be relayed can be just a set of labels that has to be converted via a look up table at the receiving end to the actual event transmitted.

The maximum entropy principle implies that for maximal code capacity, the relation between the probability of activity and its value is given by

$$p_i = \frac{exp(-\mu f_i)}{\sum_j exp(-\mu f_j)}, \tag{3}$$

where μ is interpreted as temperature in statistical mechanics formulation. Under these probabilities, the entropy of that specific code is given by

$$H_f = -\mu \sum_i p_i f_i - Z, \tag{4}$$

where Z is the partition function. If instead of using the optimal probability distribution given by (3), we use a suboptimal distribution given by q_i's, then the entropy of the new code is less or equal that given in (4), more precisely [24]

$$H_q = -\mu \sum_i q_i f_i + \sum_i q_i \log(q_i/p_i), \tag{5}$$

namely the two entropies differ by the Kullback-Leibler divergence [20] between the given distribution q and the optimal one p. The latter term is non-negative and is zero if and only if $q \equiv p$. Thus, better distributions will have a small K-L divergence and smaller differences between the distributions are more desirable.

This formulation lets us compare between different coding schemes which have the same mean activity. It can be extended to networks of neurons, and it follows that a network of BCM neurons as described in [15] maximizes entropy under the additional (independent) constraint of sparse coding.

7 Summary

Motivated by Barlow's seminal work, we have presented a theory that includes feature detection and efficient feature coding. One possible application is a fundamental neuronal task of suspicious coincidence detection. In addition, we have shown a mechanism for probability regularization for the feature detectors, so that they do not become tuned to events occurring with too small probability. The principle of maximum entropy was used to demonstrate the optimality of such neuronal code.

Acknowledgements

Fruitful discussions with Shimon Edelman, Ömer Artun and other members of the Institute for Brain and Neural Systems at Brown University are greatfully acknowledged. This work was partially supported by the Office of Naval Research.

References

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

[2] H. B. Barlow. Possible principles underlying the transfomations of sensory messages. In W. Rosenblith, editor, *Sensory Communication*, pages 217–234. MIT Press, Cambridge, MA, 1961.

[3] H. B. Barlow. Cerebral cortex as model builder. In D. Rose and V. G. Dobson, editors, *Models of the visual cortex*, pages 37–46. Wiley, New York, 1985.

[4] H. B. Barlow. Single units and sensation. *Perception*, 1:371–394, 1989.

[5] H. B. Barlow. Unsupervised learning. *Neural Computation*, 1(3):295–311, 1989.

[6] H. B. Barlow. Conditions for versatile learning, helmholtz's unconscious inference, and the task of perception. *Vision Research*, 30:1561–1571, 1990.

[7] H. B. Barlow. What is the computational goal of the neocortex. In C. Koch and J. L. Davis, editors, *Large Scale Neuronal Theories of the Brain*. MIT Press, 1994.

[8] E. L. Bienenstock, L. N Cooper, and P. W. Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal Neuroscience*, 2:32–48, 1982.

[9] S. Edelman. Class similarity and viewpoint invariance in the recognition of 3D objects. CS-TR 92-17, Weizmann Institute of Science, 1992.

[10] D. J. Field. What is the goal of sensory coding. *Neural Computation*, 6:559–601, 1994.

[11] P. Földiák. Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics*, 64:165–170, 1990.

[12] C. Fyfe and R. Baddeley. Finding compact and sparse-distributed representations of visual images. *Network*, 6:333–344, 1995.

[13] S. Geman and E. Bienenstock. Compositional vision, 1995. Talk given at the Object Features for Visual Shape Representation workshop, NIPS.

[14] N. Intrator. Feature extraction using an unsupervised neural network. *Neural Computation*, 4:98–107, 1992.

[15] N. Intrator and L. N Cooper. Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*, 5:3–17, 1992.

[16] N. Intrator and J. I. Gold. Three-dimensional object recognition of gray level images: The usefulness of distinguishing features. *Neural Computation*, 5:61–74, 1993.

[17] N. Intrator, D. Reisfeld, and Y. Yeshurun. Face recognition using a hybrid supervised/unsupervised neural network. *Pattern Recognition Letters*, 17:67–76, 1996.

[18] E. T. Jaynes. Information theory and statistical mechanics I. *Phys. Rev.*, 106:620–530, 1957.

[19] E. T. Jaynes. On the rationale of maximum entropy methods. *Proc. IEEE*, 70:939–952, 1982.

[20] S. Kullback. *Information Theory and Statistics*. John Wiley, New York, 1959.

[21] B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Network (to appear)*, 1996.

[22] G. Sebestyen. *Decision Making Processes in Pattern Recognition*. Macmillan, New York, 1962.

[23] O. G. Selfridge. Pandemonium: a paradigm for learning. In *The mechanisation of thought processes*. H.M.S.O., London, 1959.

[24] C. J. Thompson. *Classical Equilibrium Statistical Mechanics*. Clarendon Press, Oxford, 1988.

[25] Y. Weiss and S. Edelman. Representation of similarity as a goal of early visual processing. *Network*, 6:19–41, 1995.