

## The Limits of Counting Accuracy in Distributed Neural Representations

**A. R. Gardner-Medwin**

*Department of Physiology, University College London, London WC1E 6BT, U.K.*

**H. B. Barlow**

*Physiological Laboratory, Cambridge CB2 3EG, U.K.*

Learning about a causal or statistical association depends on comparing frequencies of joint occurrence with frequencies expected from separate occurrences, and to do this, events must somehow be counted. Physiological mechanisms can easily generate the necessary measures if there is a direct, one-to-one relationship between significant events and neural activity, but if the events are represented across cell populations in a distributed manner, the counting of one event will be interfered with by the occurrence of others. Although the mean interference can be allowed for, there is inevitably an increase in the variance of frequency estimates that results in the need for extra data to achieve reliable learning. This lowering of statistical efficiency (Fisher, 1925) is calculated as the ratio of the minimum to actual variance of the estimates. We define two neural models, based on presynaptic and Hebbian synaptic modification, and explore the effects of sparse coding and the relative frequencies of events on the efficiency of frequency estimates. High counting efficiency must be a desirable feature of biological representations, but the results show that the number of events that can be counted simultaneously with 50% efficiency is fewer than the number of cells or 0.1–0.25 of the number of synapses (on the two models)—many fewer than can be unambiguously represented. Direct representations would lead to greater counting efficiency, but distributed representations have the versatility of detecting and counting many unforeseen or rare events. Efficient counting of rare but important events requires that they engage more active cells than common or unimportant ones. The results suggest reasons that representations in the cerebral cortex appear to use extravagant numbers of cells and modular organization, and they emphasize the importance of neuronal trigger features and the phenomena of habituation and attention.

### 1 Introduction ---

The world we live in is highly structured, and to compete in it successfully, an animal has to be able to use the predictive power that this structure makes

possible. Evolution has molded innate genetic mechanisms that help with the universal basics of finding food, avoiding predators, selecting habitats, and so forth, but much of the structure is local, transient, and stochastic rather than universal and fully deterministic. Higher animals greatly improve the accuracy of their predictions by learning about this statistical structure through experience: they learn what sensory experiences are associated with rewards and punishments, and they also learn about contingencies and relationships between sensory experiences even when these are not directly reinforced.

Sensory inputs are graded in character and may provide weak or strong evidence for identification of a discrete binary state of the environment such as the presence or absence of a specific object. Such classifications are the data on which much simple inference is built and about which associations must be learned. Learning any association requires a quantitative step in which the frequency of a joint event is observed to be very different from the frequency predicted from the probabilities of its constituents. Without this step, associations cannot be reliably recognized, and inappropriate behavior could result from attaching too much importance to chance conjunctions or too little to genuine causal ones. Estimating a frequency depends in its turn on counting, using that word in the rather general sense of marking when a discrete event occurs and forming a measure of how many times it has occurred during some epoch.

Counting is thus a crucial prerequisite for all learning, but the form in which sensory experiences are represented limits how accurately it can be done. If there is at least one cell in a representation of the external world that fires in one-to-one relation to the occurrence of an event (i.e., if that event is directly represented according to our definitions—see Table 1), then there is no difficulty in seeing how physiological mechanisms within such a cell could generate an accurate measure of the event frequency. On the other hand, there is a problem when the events correspond to patterns on a set of neurons (i.e., with a distributed representation; see Table 1). In a distributed representation, a particular event causes a pattern of activity in several cells, but even when this pattern is unique, there is no unique element in the system that signals when the particular event occurs and does not signal at other times. Each cell active in any pattern is likely to be active for several different events during a counting epoch, so no part of the system is reliably active when, and only when, the particular event occurs.

The interference that results from this overlap in distributed representations can be dealt with in two ways: (1) cells and connections can be devoted to identifying directly in a one-to-one manner when the patterns occur (i.e., a direct representation can be generated), or (2) the interference can be accepted and the frequency of occurrence of the distributed patterns estimated from the frequency of use of their individual active elements. The second procedure is likely to increase the variance of estimated counts, and distributed representation would be disadvantageous when

Table 1: Definitions.

---

The *network* is a set of  $Z$  binary cells. An *event*, relative to the network, is any stimulation that causes a specific activity vector (pattern of  $W$  active and  $Z-W$  inactive cells). This vector is the *representation* of the event. Repeated occurrence of a representation implies repetition of the same event, even if the external stimulus is different.

A *direct* representation of an event contains at least one active cell that is active in no other event. The cells that directly represent an event in this way have a one-to-one relation between their activity and occurrences of the event.

All other representations are *distributed* representations. Each active cell is active also in other events, and identification of an event requires interaction with more than one cell. *Compact* distributed representations employ relatively few cells to distinguish a given number ( $N$ ) of events, close to the minimum  $Z = \log_2(N)$ .

The *activity ratio* of a representation is the fraction of cells ( $\alpha = W/Z$ ) that are active in it. A *sparse* representation has a low activity ratio. Direct representations are not necessarily sparse, though they must have extreme sparseness ( $W = 1$ ) to represent the greatest possible number of distinct events ( $Z$ ) on a network.

Overlap between two representations is the number of shared active cells ( $U$ ).

*Counting* of an event means estimating how many times it has occurred during a *counting epoch*. Counting accuracy is limited by overlap and by the *interference ratio* (see equation 4.2), which in simple cases is the total number of occurrences of *interfering events* (i.e., events different from the counted event) divided by occurrences of the counted event.

---

this happens because the speed and reliability of learning would be impaired.

On the other hand, distributed representation is often regarded as a desirable feature of the brain because it brings the capacity to distinguish large numbers of events with relatively few cells (see, for instance, Hinton & Anderson, 1981; Rumelhart & McClelland, 1986; Hinton, McClelland, & Rumelhart, 1986; Churchland, 1986; Farah, 1994). With sparse distributed representations, networks can also operate as content-addressable memories that store and retrieve amounts of information approaching the maximum permitted by their numbers of modifiable elements (Willshaw, Buneman, & Longuet-Higgins, 1969; Gardner-Medwin, 1976).

Recently Page (2000) has emphasized some disadvantages of distributed representations and argued that connectionist models should include a "localist" component, but we are not aware of any detailed discussion of the potential loss of counting accuracy that results from overlap, so our goal in this article is to analyze this quantitatively. To give the analysis concrete meaning, we formulated two specific neural models of the way frequency estimates could be made. Neither is intended as a direct model of the way the brain actually counts, nor do we claim that counting is the sole function of any part of the brain, but the models help to identify issues that relate more to the nature of representations than to specific mechanisms. Counting

is a necessary part of learning, and representations that could not support efficient counting could not support efficient learning.

We express our results in terms of the reduction in statistical efficiency (Fisher, 1925) of these models, since this reveals the practical consequences of the loss of counting accuracy in terms of the need for more experience before an association or contingency can be learned reliably. We do not know of any experimental measures of the statistical efficiency of a learning task, but it has a long history in sensory and perceptual psychology where, for biologically important tasks, the efficiencies are often surprisingly high (Rose, 1942; Tanner & Birdsall, 1958; Jones, 1959; Barlow, 1962; Barlow & Reeves, 1979; Barlow & Tripathy, 1997).

From our analysis we conclude that compact distributed representations (i.e., ones with little redundancy) enormously reduce the efficiency of counting and must therefore slow reliable learning, but this is not the case if they are redundant, having many more cells than are required simply for representation. The analysis enables us to identify principles for sustaining high efficiency in distributed representations, and we have confirmed some of the calculations through simulation. We think these results throw light on the complementary advantages of distributed and direct representation.

**1.1 The Statistical Efficiency of Counting.** The events we experience are often determined by chance, and it is their probabilities that matter for the determination of optimal behavior. Probability estimates must be based on finite samples of events, with inherent variability, and accurate counting is advantageous insofar as it helps to make the most efficient use of such samples. For simplicity, we analyze the common situation in which the numbers of events follow (at least approximately) a Poisson distribution about the mean, or expected, value. The variance is then equal to the mean ( $\mu$ ), and the coefficient of variation (i.e., standard deviation divided by mean) is  $1/\sqrt{\mu}$ .

A good algorithm for counting is unbiased; on average it gives the actual number within the sample, but it may nevertheless introduce a variance  $V$ . (See Table 2 for a listing of the symbols used in this article.) This variance arises within the nervous system, in a manner quite distinct from the Poisson variance whose origin is in the environment; we assume they are independent and therefore sum to a total variance  $V + \mu$ . It is convenient to consider the fractional increase of variance, caused by the algorithm in a particular context, which we call the *relative variance* ( $\rho$ ):

$$\rho = V/\mu. \quad (1.1)$$

Adding variance to a probability estimate has a similar effect to making do with a smaller sample, with a larger coefficient of variation. Following Fisher (1925) we define efficiency  $e$  as the fraction of the sample that is effectively

Table 2: Principal Symbols Employed.

$Z$	number of binary cells in the network
$P_i$	a binary pattern (or vector) of activity on the network
$E_i$	an event causing the pattern $P_i$ (its representation)
$N$	the number of such distinct events that may occur with finite probability during a counting epoch
$W_i$	number of active cells in the representation of $E_i$
$\alpha_i = W_i/Z$	activity ratio for the representation of $E_i$
$U_{ij}$	overlap (i.e., number of shared active cells) between $P_i, P_j$
$\mu_i$	expectation of the number of occurrences of $E_i$ in a counting epoch
$m_i$	actual number of occurrences of $E_i$
$M = \sum_i m_i$	total number of event occurrences within the counting epoch
$V$	variance introduced in estimating a count $m$
$\rho = V/\mu$	relative variance, that is, the variance of the estimate of $m$ relative to the intrinsic Poisson variance of $m (= \mu)$
$e = \mu/(\mu + V)$	efficiency in estimating $\mu$ , given the variance $V$ in counting a sample
$\Phi_c$	interference ratio while counting $E_c$ (see equation 4.2)
$\langle y \rangle$	the statistical expectation of any variable $y$
$\{y_i\}$	the set of $y_i$ for all possible values of $i$
$\hat{y}$	an estimate of $y$

made use of:

$$e = \mu/(\mu + V) = (1 + \rho)^{-1}. \quad (1.2)$$

Efficiency is a direct function of  $\rho$ , and if  $\rho > 1$ , then  $e < 50\%$ , which means that the time and resources required to gather reliable data will be more than two times greater than is in principle achievable with an ideal algorithm. If  $\rho \ll 1$ , then efficiency is nearly 100% and there would be little to gain from a better algorithm in the same situation.

## 2 A Simple Illustration

As an illustration of the problem, consider how to count the occurrences of a particular letter (e.g.,  $A$ ) in a sequence of letters forming some text. If  $A$  has a direct representation in the sense that an element is active when and only when  $A$  occurs (as on a keyboard), then it is easy to count the occurrences of  $A$  with precision. But if  $A$  is represented by a distinctive pattern of active elements (as in the ASCII code), then the problem is to infer from counts of usage of individual elements whether and how often  $A$  has occurred. The ASCII code is compact, with 127 keyboard and control characters distinguished on just 7 bits. Obviously 7 usage counts cannot in

general provide enough information to infer 127 different counts precisely. The result is underdetermined except for a few simple cases. In general there is only a statistical relation between the occurrence of letters and the usage of their representational elements, and our problem is to calculate, for cases ranging from direct representations to compact codes, how much variance is added when inferring these frequencies.

Note that seven specific subsets of characters have a one-to-one relation to activity on individual bits in the code. For example, the least significant bit is active for a set including the characters *ACEG* as well as many others. Such subsets have special significance because the summed occurrences of events in them are easily computed on the corresponding bit. In the ASCII code, they are generally not subsets of particular interest, but in the brain, it would be advantageous for them to correspond to categories of events that can be grouped together for learning. This would improve generalization, increase the sample size for learning about the categories, and reduce the consequences of overlap errors. Our analysis ignores the benefit from such organization and assumes that the representations of different events are randomly related, though we discuss this further in section 6.2.

The conversion of directly represented key presses into a distributed ASCII code is certainly not advantageous for the purpose of counting characters. The events that the brain must count, however, are not often directly represented at an early stage, nor do they occur one at a time in a mutually exclusive manner, as do typed characters. Each event may arouse widespread and varied activity that requires much neural computation before it is organized in a consistent form, suitable for counting and learning. We assume here that perceptual mechanisms exploit the redundancy of sensory messages and generate suitable inputs for our models as outlined below and discussed later (in section 6). These simplifications enable us to focus on the limitations of counting accuracy that arise even under relatively ideal conditions.

### 3 Formal Definition of the Task

---

Consider a set of  $Z$  binary cells (see Figure 1) on which is generated, one at a time, a sequence of patterns of activity belonging to a set  $\{P_1, \dots, P_N\}$  that correspond to  $N$  distinct categorizations of the environment described as events  $\{E_1, \dots, E_N\}$ . The patterns (binary vectors) are said to represent the events. Each pattern  $P_i$  is an independent random selection of  $W_i$  active cells out of the  $Z$  cells, with the activity ratio  $\alpha_i = W_i/Z$ . The corresponding vector  $\{x_{i1}, \dots, x_{iZ}\}$  has elements 1 or 0 where cells are active or inactive in  $P_i$ . The active cells in two different patterns  $P_i, P_j$  overlap by  $U_{ij}$  cells ( $U_{ij} \geq 0$ ), where

$$U_{ij} = \sum_{k=1, Z} (x_{ik}x_{jk}).$$

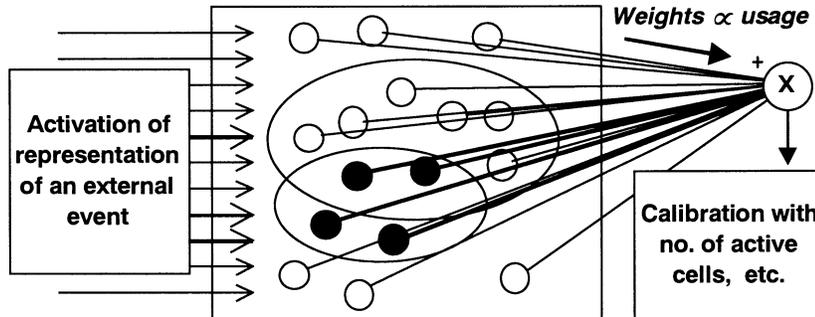


Figure 1: Outline of the projection model. Sets of binary cells (for example, those marked by ellipses) are activated when there are repeatable sensory or neural events. The frequency of occurrence of a particular event  $E_c$  (with its active cells black) is estimated from activation of an accumulator cell  $X$  when  $E_c$  is represented at the time of testing. Synaptic weights onto  $X$  are proportional to the usage of individual cells. Additional circuitry (not shown) counts the number of active cells  $W_c$  and estimates both the average number of active cells and the total number of events ( $M$ ) during the counting period.

Note that two different events may occasionally have identical representations, since these are assigned at random.

Consider an epoch during which the total number of occurrences  $\{m_i\}$  of events  $\{E_i\}$  can be treated as independent Poisson variables with expectations  $\{\mu_i\}$ . The totals  $M$  and  $\mu_T$  are defined as  $M = \sum_i(m_i)$  and  $\mu_T = \sum_i(\mu_i)$ . The task we define is to estimate, using only plausible neural mechanisms, the actual number of occurrences ( $m_c$ ) of representations of an individual event  $E_c$  when this event is identified by a test presentation after the counting epoch. We suppose that the system can employ an accurate count of the total occurrences ( $M$ ) summed over all events during the epoch, and also the average activity ratio  $\bar{\alpha}$  during the epoch:

$$\bar{\alpha} = \sum_i(m_i \alpha_i) / M. \quad (3.1)$$

We require specific and neurally plausible models of the way the task is done, and these are described in the next two sections. The first model (in section 3.1) is based on modifiable projections from the cells of a representation. They support counting by increasing effectiveness in proportion to presynaptic usage, though associative changes or habituation might alternatively support learning or novelty detection with similar constraints. The second model (in section 3.2) is based on changes of internal support for a pattern of activation. This greatly adds to the number of variables available to store relevant information by involving modifiable synapses between ele-

ments of a distributed representation, analogous to the rich interconnections of the cerebral cortex.

Readers wishing to skip the mathematical derivations in sections 3.1 and 3.2 should look at their initial paragraphs with Figures 1 and 2 and proceed to section 4.

**3.1 The Projection Model.** This model (see Figure 1) estimates  $m_c$  by obtaining a sum  $S_c$  of the usage, during the epoch, of all those cells that are active in the representation of  $E_c$ . This computation is readily carried out with a single accumulator cell  $X$  (see Figure 1) onto which project synapses from all the  $Z$  cells.

The strengths of these synapses increase in proportion to their usage. When the event  $E_c$  is presented in a test situation after the end of an epoch of such counting, the summed activation onto  $X$  gives the desired total:

$$\begin{aligned} S_c &= \sum_{\text{cells } k} \left( x_{ck} \sum_{\text{events } j} (x_{jk} m_j) \right) \\ &= m_c W_c + \sum_{\text{events } j \neq c} (m_j U_{jc}). \end{aligned} \quad (3.2)$$

If there were no interference from overlap between active cells in  $E_c$  and in any other events occurring during the epoch (i.e., if  $m_j = 0$  for all  $j$  for which  $U_{jc} > 0$ ), then  $S_c = m_c W_c$ . In this situation,  $S_c/W_c$  gives a precise estimate of  $m_c$  and is easily computed since  $W_c$  is the number of cells active during the test presentation of  $E_c$ . In general,  $S_c$  will be larger than  $m_c W_c$  due to overlap between event representations. An adjustment for this can be made on the basis of the total number of events  $M$  and the average activity ratio  $\bar{\alpha}$ , yielding a revised sum  $S'_c$ :

$$S'_c = S_c - M W_c \bar{\alpha}. \quad (3.3)$$

Expansion using equations 3.1 and 3.2 yields:

$$S'_c = m_c W_c (1 - \alpha_c) + \sum_{\text{events } j \neq c} (m_j (U_{jc} - \alpha_j W_c)). \quad (3.4)$$

The expectation of each term in the sum in equation 3.4 is zero, since  $\langle U_{jc} \rangle = \alpha_j W_c$  and the covariance for variations of  $m_j$  and  $U_{jc}$  is zero since they are determined by independent processes. An unbiased estimate  $\hat{m}_c$  of  $m_c$  is therefore given by:

$$\hat{m}_c = S'_c / (W_c (1 - \alpha_c)). \quad (3.5)$$

To calculate the reliability and statistical efficiency of this estimate  $\hat{m}_c$ , we need to know the variance of  $S'_c$  due to the interference terms in equation 3.4.

This is simplified by the facts that  $m_j$  and  $U_{jc}$  vary independently and that  $\langle U_{jc} \rangle = \alpha_j W_c$ :

$$\text{Var}(S'_c) = \sum_{j \neq c} (\langle m_j \rangle^2 \text{Var}(U_{jc}) + \text{Var}(U_{jc}) \text{Var}(m_j)). \quad (3.6)$$

$U_{jc}$  has a hypergeometric distribution, close to a binomial distribution, with expectation  $\alpha_j W_c$  and variance  $\alpha_j(1 - \alpha_j)(1 - \alpha_c)(1 - 1/Z)^{-1} W_c$ . Substituting these values and  $\langle m_j \rangle = \text{Var}(m_j) = \mu_j$  for the Poisson distribution of  $m_j$  we obtain:

$$\text{Var}(S'_c) = W_c \sum_{j \neq c} (\alpha_j(1 - \alpha_j)(1 - \alpha_c)(1 - 1/Z)^{-1} (\mu_j + \mu_j^2)). \quad (3.7)$$

Note that this analysis includes two sources of uncertainty in estimates of  $m_c$ : variation of the frequencies of interfering events around their means and uncertainty of the overlap of representations of individual interfering events. The overlaps between different representations are strictly fixed quantities for a given nervous system, so this source of variation does not contribute if the nervous system can adapt appropriately. Results of calculations are therefore given for both the full variance (see equation 3.7) and the expectation value of  $\langle \text{Var}_m(S'_c) \rangle$  when  $\{U_{jc}\}$  is fixed—that is, for variations of  $\{m_j\}$  alone. This modified result is obtained as follows. Instead of equation 3.6, we have the variance of  $S'_c$  (see equation 3.4) due to variations of  $\{m_j\}$  alone:

$$\text{Var}_m(S'_c) = \sum_{j \neq c} ((U_{jc} - \alpha_j W_c)^2 \mu_j). \quad (3.8)$$

This depends on the particular (fixed) values of  $\{U_{jc}\}$ , but we can calculate its expectation for randomly selected representations:

$$\begin{aligned} \langle \text{Var}_m(S'_c) \rangle &= \sum_{j \neq c} (\langle (U_{jc}^2) \rangle - 2\alpha_j W_c \langle U_{jc} \rangle + \alpha_j^2 W_c^2) \mu_j \\ &= W_c \sum_{j \neq c} (\alpha_j(1 - \alpha_j)(1 - \alpha_c)(1 - 1/Z) \mu_j). \end{aligned} \quad (3.9)$$

This expression is similar to equation 3.7, omitting the terms  $\mu_j^2$ . Note that if  $\mu_j \ll 1$  for all events that might occur, the difference between the two expressions is negligible. This corresponds to a situation where there may be many possible interfering events, but each one has a low probability of occurrence. The variance does not then depend on whether individual overlaps are fixed and known, since the events that occur are themselves unpredictable.

The relative variance  $\rho$  (see equation 1.1) for an estimate  $\hat{m}_c$  of a desired count is obtained by dividing the results in equations 3.7 and 3.9 first by the square of the divisor in equation 3.5 ( $W_c(1-\alpha_c)^2$ ) and then by the expectation of the count  $\mu_c$ . Square brackets are used to denote the terms in equation 3.7 due to overlap variance that are omitted in equation 3.9:

$$\rho(\hat{m}_c) = \frac{1}{(Z-1)} \times \frac{\sum_{j \neq c} (\alpha_j(1-\alpha_j)(\mu_j + \mu_j^2))}{\alpha_c(1-\alpha_c)\mu_c}. \quad (3.10)$$

**3.2 The Internal Support Model.** In the projection model, the stored variables correspond to the usage of individual cells. The number of such variables is restricted to the number of cells  $Z$ , and this is a limiting factor in inferring event frequencies. The second model takes advantage of the greater number of synapses connecting cells, which in the cerebral cortex can exceed the number of cells by factors of  $10^3$  or more. The numbers of pairings of pre- and postsynaptic activity can be stored by Hebbian mechanisms and yield substantially independent information at separate synapses (Gardner-Medwin, 1969). The number of stored variables for a task analogous to counting is greatly increased, though at the cost of more complex mechanisms for handling the information.

The support model (see Figure 2) employs excitatory synapses that acquire strengths proportional to the number of times the pre- and postsynaptic cells have been active together during events experienced in a counting epoch. Full connectivity ( $Z^2$  synapses counting all possible pairings of activity) is assumed here in order to establish optimum performance, though the same principles would apply in a degraded manner with sparse connectivity. During test presentation of the counted event  $E_c$ , the potentiated synapses between its active cells act in an autoassociative manner to give excitatory support for maintenance of the representation (Gardner-Medwin, 1976, 1989). The extent of this internal excitatory support depends substantially on whether, and how often,  $E_c$  has been active during the epoch. Interference is caused by overlapping events, just as with the projection model, though with less impact because the shared fraction of pairings of cell activity is, with sparse representations, much less than the shared fraction of active cells.

Measurement of internally generated excitation requires appropriate external handling of the cell population (see Figure 2). In principle, the whole histogram of internal excitation onto different cells can be established by imposing fluctuating levels of diffuse inhibition along with activation of the pattern representing  $E_c$ . Our analysis requires just the total (or average) excitation onto the cells of  $E_c$  (see Equation 3.11). The neural dynamics may introduce practical limitations on the accuracy of such a measure in a given period of time, so our results represent an upper bound on performance employing this model.

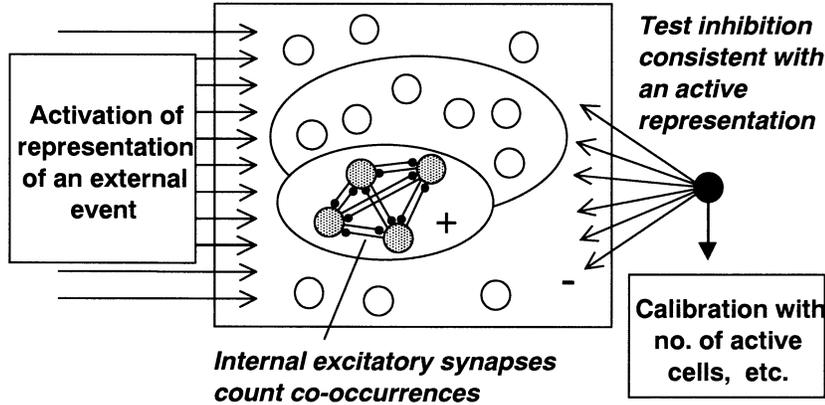


Figure 2: Outline of the internal support model. Internal excitatory synapses within the network measure the frequency of co-occurrences of activity in pairs of cells by a Hebbian mechanism. On a test representation of the event  $E_c$  to be counted (indicated by the hatched active cells), the total of the internal activation stabilizing the active pattern is estimated by testing the effect of a diffuse inhibitory influence on the number of active cells.

We restrict the analysis for simplicity to situations with equal activity ratios for all events ( $\alpha_j \equiv \alpha$ ;  $W_j \equiv W$ ) and full connectivity. Each of  $Z$  cells is connected to every other cell with Hebbian synapses counting pre- and postsynaptic coincidences, including autapses that effectively count individual cell usage. Analysis follows the same lines as for the projection model (see section 3.1), and only the principal results are stated here, with some steps omitted.

When  $E_c$  is presented as a test stimulus, the total excitation  $Q_c$  from one cell to another within  $E_c$  is given, analogous to equation 3.2, by:

$$Q_c = m_c W^2 + \sum_{\text{events } j \neq c} (m_j U_{jc}^2). \quad (3.11)$$

A corrected value  $Q'_c$  is calculated to allow for average levels of interference:

$$Q'_c = Q_c - M\alpha W(\alpha W + 1 - 2\alpha)(1 - 1/Z)^{-1}. \quad (3.12)$$

This has expectation equal to  $m_c W^2(1 - \alpha)(1 + \alpha - 2/Z)(1 - 1/Z)^{-1}$  so we can obtain an unbiased estimate of  $m_c$  as follows:

$$\hat{m}_c = Q'_c / (W^2(1 - \alpha)(1 + \alpha - 2/Z)(1 - 1/Z)^{-1}). \quad (3.13)$$

The full variance of  $Q'_c$  is calculated as for  $S'_c$  (see equation 3.6), taking account of the independent Poisson distributions for the numbers of interfering events  $\{m_j\}$  and hypergeometric distributions for the overlaps  $\{U_{jc}\}$ :

$$\text{Var}(Q'_c) = \text{Var}(U_{jc}^2) \sum_{j \neq c} (\mu_j + \mu_j^2). \quad (3.14)$$

The algebraic expansion of  $\text{Var}(U_{jc}^2)$  is complex, but is simplified with the terminology  $F_{(r)} = (W!/(W-r)!)^2(Z-r)!/Z!$ :

$$\text{Var}(U_{jc}^2) = (F_{(4)} + 6F_{(3)} + 7F_{(2)} + F_{(1)} - (F_{(2)} + F_{(1)})^2). \quad (3.15)$$

The relative variance for the frequency estimate  $\hat{m}_c$  (see equation 3.13) can then be written as:

$$\rho(\hat{m}_c) = \frac{\xi}{Z^2} \times \frac{\mu_I}{\mu_c} [ \times (1 + \bar{\mu}_i) ], \quad (3.16)$$

where  $\mu_I = \sum_{j \neq c} \mu_j$  is the expected total number of occurrences of interfering events,  $\bar{\mu}_i = \sum_{j \neq c} \mu_j^2 / \mu_I$  is the average number of repeats of individual interfering events, weighted according to their frequencies, and  $\xi$  is given by:

$$\xi = \frac{F_{(4)} + 6F_{(3)} + 7F_{(2)} + F_{(1)} - (F_{(2)} + F_{(1)})^2}{W^4 Z^{-2} (1 - W/Z)^2 (1 + (W-2)/Z)^2 (1 - 1/Z)^{-2}}. \quad (3.17)$$

$\xi$  depends on both  $W$  and  $Z$ , but for networks of different size ( $Z$ ), it has a minimum value for an optimal choice of  $W$  ( $\approx \sqrt{Z/2}$ ) that is only weakly dependent on  $Z$ :  $\xi_{\min} = 3.9$  for  $Z = 10$ , 6.7 for  $Z = 100$ , 8.7 for  $Z = 1000$ , 9.9 for  $Z = 10^5$ . The corresponding optimal activity ratio, to give minimum variance and maximum counting efficiency with this model, is therefore  $\alpha \approx 1/\sqrt{2Z}$ .

#### 4 Results for Events Represented with Equal Activity Ratios

Analysis for the support model was restricted, for simplicity, to cases where all activity ratios are equal. In this section we also apply this restriction to the projection model to assist initial interpretation and comparisons. The relative variance for the projection model (see equation 3.10) becomes:

$$\rho(\hat{m}_c) = \frac{1}{(Z-1)} \times \frac{\mu_I}{\mu_c} [ \times (1 + \bar{\mu}_i) ]. \quad (4.1)$$

Both this and the corresponding equation 3.16 for the support model can be broken down as products of a representation-dependent term  $((Z-1)^{-1}$

or  $\xi Z^{-2}$ ) and a term that depends on only the expected frequencies of the counted and interfering events. The latter term is the same for both models, and we call it the *interference ratio* ( $\Phi_c$ ) for a particular event  $E_c$ :

$$\Phi_c = \frac{\text{Expected occurrences of events other than the counted event}}{\text{Expected occurrences of counted event}} [\times (1 + \bar{\mu}_i)]. \quad (4.2)$$

The interference ratio expresses the extent to which a counted event is likely to be swamped by interfering events during the counting epoch. The principal determinant is the ratio of occurrences of interfering and counted events:  $\mu_I/\mu_c$ . The term in square brackets expresses the added uncertainty that can be introduced when interfering events occur with multiple repetitions ( $\bar{\mu}_i \gg 1$ ), because overlaps that are above and below expectation do not then average out so effectively. As described after equation 3.7, stable conditions may allow the nervous system to adapt to compensate for a fixed set of overlaps, corresponding to omission of this term; it is in any case negligible if interfering events seldom repeat ( $\bar{\mu}_i \ll 1$ ).

We can see how  $\Phi_c$  governs the relative variance of the count by substituting in equations 4.1 and 3.16 for the two models:

$$\rho_{\text{projection}}(\hat{m}_c) = \frac{\Phi_c}{Z - 1} \quad (4.3)$$

$$\rho_{\text{support}}(\hat{m}_c) = \xi \times \frac{\Phi_c}{Z^2}. \quad (4.4)$$

If the number of cells (or synapses, for the support model) is much less than the interference ratio ( $Z \ll \Phi_c$  or  $Z^2 \ll \xi \Phi_c$ ), then  $\rho \gg 1$  and counting efficiency is necessarily very low. High efficiency ( $> 50\%$ ) requires  $\rho < 1$ , and networks that are correspondingly much larger and very redundant from an information-theoretic point of view (see section 6.1). For a particular number of cells  $Z$ , 50% efficiency for an event  $E_c$  requires that  $\Phi_c < Z$  or  $Z^2/\xi$  on the two models. For the support model, the highest levels of interference are tolerated with sparse representations giving minimum  $\xi$ , with  $\alpha$  approximately  $1/\sqrt{2Z}$  and  $\Phi_c < 0.1 - 0.25Z^2$  (see section 3.2).

If we set a given criterion of efficiency, the maximum interference ratio  $\Phi_c$  scales in proportion to  $Z$  for the projection model and approximately  $Z^2$  for the support model. This is consistent with what one might expect given that the numbers of underlying variables used to accumulate counts are, respectively,  $Z$  and  $Z^2$  in the two models, though the performance per cell in the projection model is up to 10 times greater than the performance per synapse in the support model.

Figure 3 illustrates the dependence of efficiency on the number of cells  $Z$ , for an interference ratio  $\Phi_c = 100$  and  $\alpha = 0.1$ . When  $Z$  exceeds 100 ( $= \Phi_c$ ),

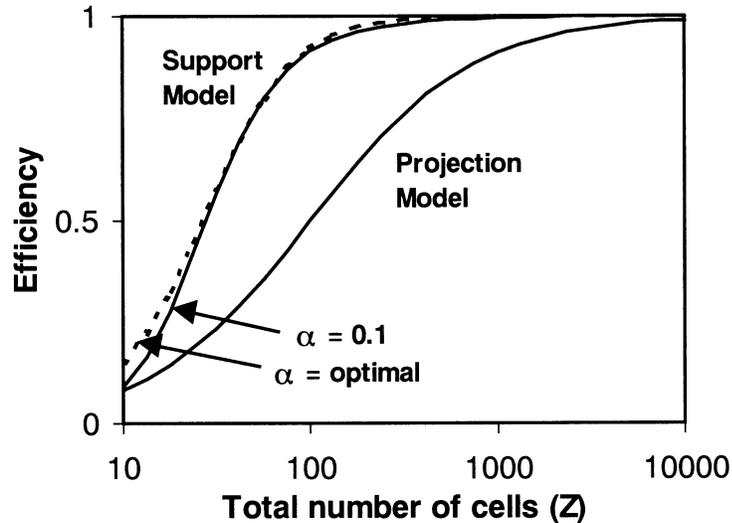


Figure 3: Counting efficiency as a function of the number of cells ( $Z$ ) used for distributed representations assumed to have the same activity ratio ( $\alpha$ ) for all events. Calculations are for  $\Phi_c = 100$  (i.e., 100 times more total occurrences of interfering events than of the counted event). The full lines are for  $\alpha = 0.1$  with the projection model (heavy line) and support model (thin line) and for  $\alpha = 1/\sqrt{2Z}$  for maximum efficiency using the support model (dotted line). Note that if all events are equiprobable, implying that there is a total of just 101 different event types occurring, then seven cells would suffice to represent them distinctly if there were no need for counting. High counting efficiency requires 10 times this number, and the corresponding ratio would be higher for larger numbers of events.

the efficiency exceeds about 50% on the projection model and 93% on the internal support model. The advantage of the second model (based on counts of paired cellular activity) results from the fact that the proportion of pairings shared with a typical interfering event ( $\alpha^2$ ) is less than the proportion of shared active cells ( $\alpha$ ). Efficiency with the projection model is independent of  $\alpha$ , while with the support model, it is maximized by choosing  $\alpha$  to minimize  $\xi$  in equation 4.4 (see above).

Simulations were performed using LABVIEW (National Instruments) to confirm some of the results in the foregoing analysis. Figures 4A and 4B show simulation results for conditions for which Figure 3 gives the theoretical expectations ( $N = 101$  equiprobable events,  $\mu = 10$ ,  $\Phi_c = 100$ ,  $\alpha = 0.1$ ,  $Z = 100$ ). The observed efficiencies are in agreement (see the legend to Figure 4), and the graphs illustrate the extent of correspondence between estimated and true counts. The horizontal spread on these graphs shows

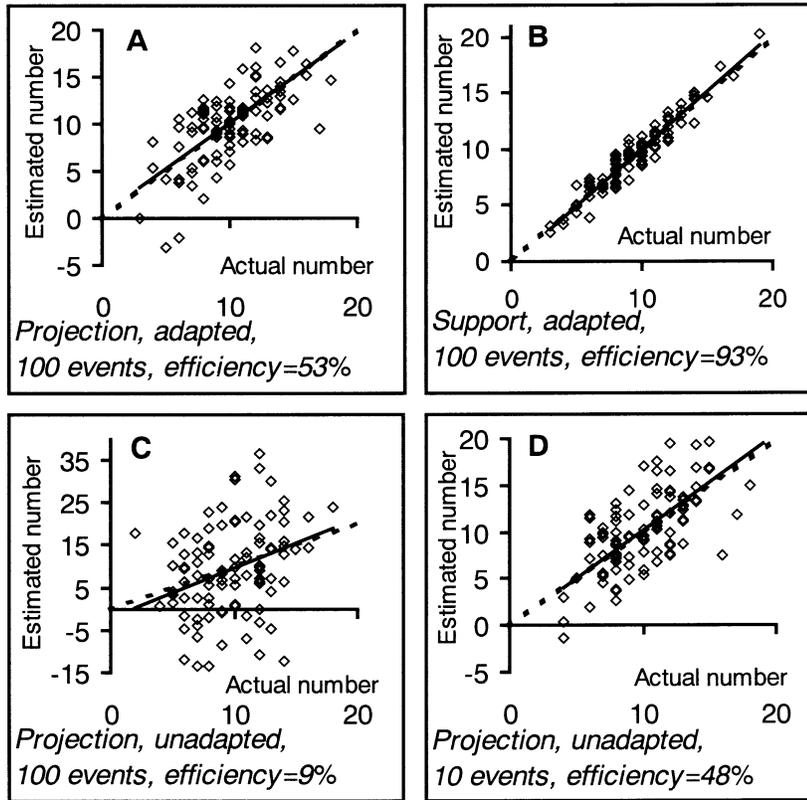


Figure 4: Estimated counts plotted against actual counts, obtained by simulation with the projection model (A,C,D) and the support model (B). The network had 100 cells ( $Z = 100$ ) with 10 active during events ( $W = 10, \alpha = 0.1$ ), selected at random for each of 101 different events for A,B,C and 10 for D. The numbers of occurrences of each event were independent Poisson variables with mean  $\mu = 10$ . New sets of representations and counts were generated to calculate each point. Estimated counts used the algorithms either with (A,B) or without (C,D) adaptation to the actual overlapping of representations of interfering events with the counted event. Interference ratios ( $\Phi_c$ : equation 4.2) were therefore 100 for A,B,D and 1089 for C. Perfect estimates would lie on the line of equality (dashed). The solid line shows the linear regression of estimates on actual counts. Efficiency  $e$  is shown, calculated as the mean squared error for the 100 points, divided by  $\mu$ . This was not significantly different (in five repeats of such simulations) from efficiencies expected from the theoretical analysis (equations 4.3, 4.4, 1.2), which were A:50%, B:93.3%, C:8.3%, D:50%.

the Poisson variation of the true counts about their mean ( $\mu = 10$ ), while the vertical deviation from the (dashed) lines of equality shows the added variance due to the algorithms. The closeness of the regression lines and the lines of equality (ideal performance) shows that the algorithms are unbiased, while the efficiency is the mean squared error divided by  $\mu$  (almost equivalent to the squared correlation coefficient  $r^2$ ). Performance is better for the support model (see Figure 4B) than the projection model (see Figure 4A) and is worse (see Figure 4C) if adaptation to fixed stochastic conditions and overlaps is not employed. The interference ratio ( $\Phi_c$ : see equation 4.2) rose for Figure 4C to  $\Phi_c = 1089$  with the same number of events handled by the network. With just 10 events (substantially fewer than the number of cells),  $\Phi_c$  was restored to 100 and the efficiency to 50%, as predicted (see Figure 4D).

The theoretical dependence of efficiency on  $\alpha$ ,  $Z$ , and  $\Phi_c$  for representations having uniform  $\alpha$  is shown in contour plots in Figures 5A and 5B for the projection and support models, respectively. Contours of equal efficiency (for  $\Phi_c = 100$ ) are plotted against  $\alpha$  and  $\log(Z/\Phi_c)$ . For the projection model (see Figure 5A) the efficiency is independent of  $\alpha$  and depends on only the ratio  $Z/\Phi_c$  (or, more strictly,  $(Z - 1)/\Phi_c$ : see equation 4.3). This graph would not be significantly different for any larger value of  $\Phi_c$ . For the support model (see Figure 5B) the efficiency is higher than for the projection model (see Figure 5A) for all combinations of  $Z/\Phi_c$  and  $\alpha$ , especially if  $\alpha$  is close to the optimum level of sparseness corresponding to the contour minima ( $\alpha \approx 1/\sqrt{2Z}$ ). Higher interference ratios ( $\Phi_c > 100$ ) yield even greater efficiencies with a given value of  $Z/\Phi_c$ , while the benefits of sparse coding with the support model become more pronounced.

Some of the implications of these plots are illustrated here with an example. Suppose initially that events are represented by 12 active cells on a population of 30 ( $\alpha = 0.4$ ,  $Z = 30$ ). With an interference ratio  $\Phi_c = 100$ , the counting efficiency would be 22% on the projection model and 45% on the support model (points marked *a* on Figures 4A and 4B). These are modest efficiencies, but changing the representations with the same event statistics can improve the efficiency.

Representing events with more sparse activity on the same number of cells, as might be achieved by simply raising activation thresholds, can improve counting efficiency with the support model, though not with the projection model. Representation with 4 instead of 12 active cells out of 30 ( $\alpha = 0.13$ ) gives efficiencies of 22%, 63% on the two models (points *b*). Though efficiency is improved with the support model, information may be lost by this encoding since there are only about  $10^4$  distinct patterns with 4 active cells out of 30, compared with  $10^8$  having 12 active. A better strategy is to use more cells. For example, recoding to 5 active cells on 100, we get 50%, 93% efficiency on the two models (points *c*), and with 4 active on 300 we get 75%, 98% (points *d*). Each of these encodings retains about  $10^8$

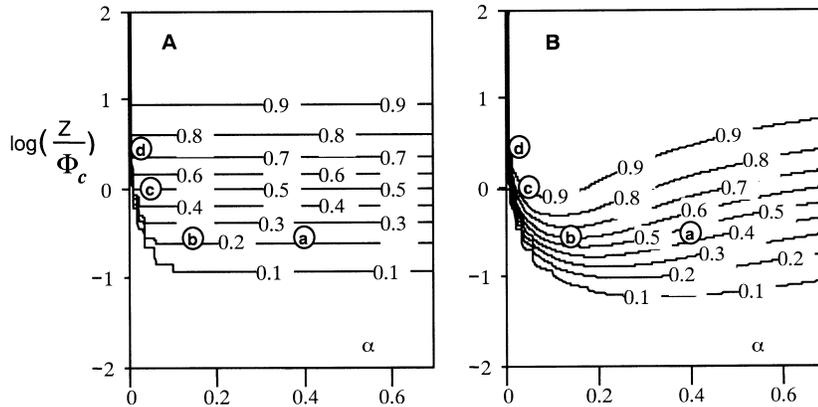


Figure 5: Plots showing counting efficiency (indicated on contours) as a function of activity ratio  $\alpha$  on the horizontal axis (assumed equal for all events) and  $\log_{10}$  of the factor by which the number of cells  $Z$  exceeds the interference ratio  $\Phi_c$  (vertical axis). Plots are for the projection model (A) and the support model (B), calculated for  $\Phi_c = 100$ , though plot A is essentially identical for all larger values of  $\Phi_c$ , apart from the cut-off at low  $\alpha$  corresponding to the requirement  $W \geq 1$ . Points *a–d* are referred to in the text.

distinguishable patterns, so incurs no loss of information. The improvement in counting efficiency is achieved at the cost of extra redundancy, with, in this example, a tenfold increase of the number of cells.

These levels of efficiency fall short of the 100% efficiency that can be achieved if cells are assigned to direct representations for each event of interest. One hundred cells would be enough (without duplication or spare capacity) to provide direct representations for 100 events, which is the largest number that can simultaneously have interference ratios  $\Phi_c = 100$ . If these 100 cells are used instead for a distributed representation for these 100 events, then there would be a moderate loss of efficiency to 50% or 93% on the two models. The merits of distributed and direct representation are considered further in the discussion, but these results suggest that if distributed representations are to be used for efficient counting, then just as many cells may be required as for direct representations. However, their ability to represent rare and novel events unambiguously gives them greater flexibility even in relation to counting.

## 5 Events with Different Probabilities and Different Activity Ratios

The interference ratio  $\Phi_c$  is higher for rare events than for common events, since the factor that varies most in proportionate terms in equation 4.2 is the denominator  $\mu_c$ . If, as assumed in the last section, all event representations

have the same activity ratio  $\alpha$ , this means that rare events (with probability less than about  $1/Z$  on the projection model) are counted inefficiently because the overlap errors act like a source of intrinsic noise. Although only relatively few events (at most  $Z$  or  $Z^2/\xi$ ) can simultaneously be counted with  $\geq 50\%$  efficiency, many of the huge number with distinct representations on a network may occur, but too rarely to be countable. Since the relative frequencies of different events are not stable features of an animal's environment, however, an infrequent event in one epoch may be counted efficiently in different epochs when it occurs more frequently (see section 6.4).

The poor counting efficiency for rare events may also be boosted if they are represented with higher activity ratios than other events. This requires that particular events be identified as worthy of such amplification, through being recognized as important, or simply novel. Conversely, lowering  $\alpha$  for common or unimportant events will be beneficial. The mechanisms that may vary activity ratios are discussed in section 6.3. The results are illustrated in Figure 6 for a simple example using the projection model, based on equation 3.10. Calculations are for 7 events with different frequencies—first with equal activity ratios  $\alpha = 0.02$  (squares) and then with activity ratios inversely proportional to frequency (crosses), giving more uniform efficiency. The way in which rare events benefit from higher  $\alpha$  is that the high probability that individual cells will be active for other interfering events is compensated by the pooling of information from more active cells. Experimentation with different power law relationships  $\alpha = k\mu^{-n}$  required  $n$  in the range 1.0 to 1.3 to give approximately uniform efficiency over a range of conditions, with  $n \sim 1.0$  when all values of  $\alpha$  are  $\ll 1$ . Although these results are presented only for the projection model, it is clear that qualitatively similar conclusions must apply for the support model.

## 6 Discussion

---

Counting events, or determining their frequencies, is necessary for learning and for appropriate control of behavior. This article analyzes the uncertainty in estimating frequencies that arises from sharing active elements between the distributed representations of different events.

To analyze this problem, we make substantial simplifying assumptions. We treat neurons as binary, with different events represented by randomly related selections of active cells. We treat the frequencies of interfering events as independent Poisson variables. And we employ just two simple models as counting algorithms, with numbers of physiological variables equal to the numbers of cells in one case and synapses in the other.

Sensory information contains a great deal of associative statistical structure that is absent from the representations we model. Our results would require modification for structured data, but it has long been argued, with considerable supporting evidence (Attneave, 1954; Barlow, 1959, 1989; Watanabe, 1960; Atick, 1992), that a prime function of sensory and perceptual pro-

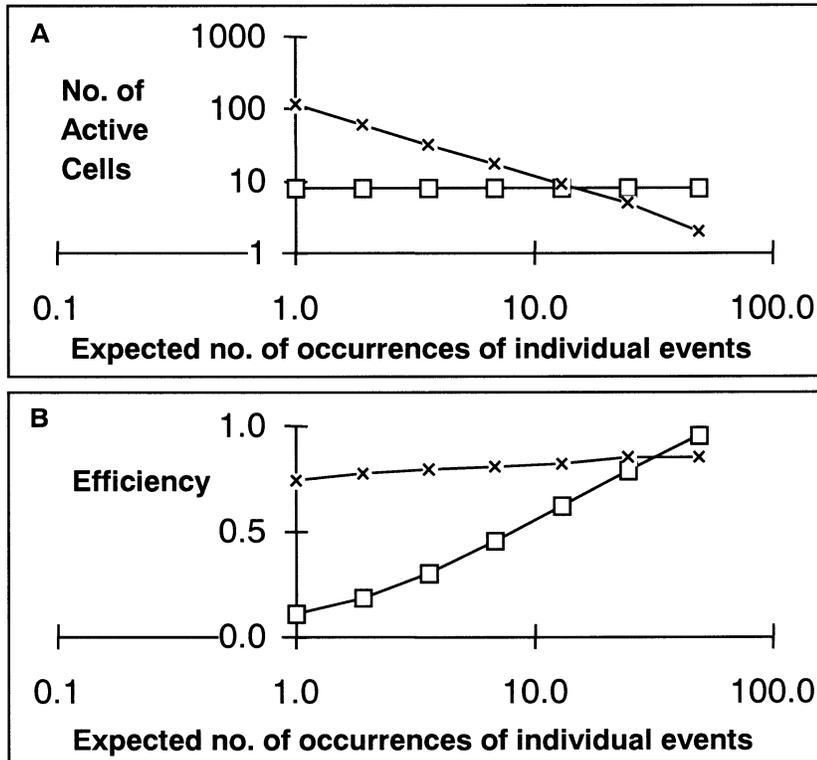


Figure 6: Effect of nonuniform event frequencies on counting efficiency. The number of active cells (A) and the counting efficiency (B) are shown for seven events with widely differing average counts during a counting epoch (horizontal axes). Two different assumptions are made: events are represented each with the same number of active cells ( $\square$ ), or with a number inversely proportional to the event probability ( $\times$ ). In each case the mean value of  $\alpha$  (weighted with probabilities of occurrence) is 0.02. Calculations are for the projection model with  $Z = 400$  cells and full variances (see equation 3.10). Rare events have lower counting efficiency when all representations have equal activity ratios, but approximately uniform efficiency is achieved ( $\times$ ) when more active cells are used to represent rare events.

cessing is to remove much of the associative statistical structure of the raw sensory input by using knowledge of what is predictable. The resulting less structured input would be closer to what we have assumed for our analysis. Notice, however, that inference-like processes are involved in unsupervised learning and perceptual coding, and that these will be influenced by errors of counting in ways similar to those analyzed here.

Our aim in the discussion is to identify simple insights and principles that are likely to apply to all processes of adaptation and learning that depend on counting. There are five sections: the counting problem in general, five principles for reducing the interference problem, possible applications of these principles in the brain, relative merits of direct and distributed representations, and the problem of handling the huge numbers of representable events on a distributed network.

**6.1 The Counting Problem in Distributed Representations.** Counting with distributed representations nearly always leads to errors, but the loss of efficiency need not be great if the representations follow principles outlined in the next section. We shall see that the capacity of a network to count different events (and therefore to learn about them) is generally enormously less than its capacity to represent different events.

The interference ratio  $\Phi_c$ , defined in equation 4.2, gives direct insight into what limits counting efficiency. At its simplest, this is the ratio of the total number of interfering events (of any type) to the number of events being counted over a defined counting epoch. In our first (projection) model, the number of cells employed in the representations ( $Z$ ) must exceed  $\Phi_c$  if the efficiency is to remain above 50%; in the second (internal support) model, fewer cells may suffice for equivalent performance, provided the number of synapses ( $Z^2$ ) is at least four to ten times  $\Phi_c$  (see equation 4.4). Increasing the number of cells increases efficiency, and with either model, three times as many cells or synapses are required to achieve 75% efficiency and nine times as many for 90% efficiency. Multiple repetitions of individual interfering events tend to amplify the errors due to overlap and raise these requirements further (see equation 4.2).

In a situation where all events are equally frequent, the interference ratio is just the number of different events, and this number is limited, as above, to roughly  $Z$  or  $Z^2/10$  for 50% efficiency. In the more general case where events have different probabilities or activate different numbers of cells, their counting efficiencies will vary; roughly speaking, it is those with the largest products of activity ratio and expected frequency ( $\alpha_i \mu_i$ ) that have the greatest counting efficiency (see section 5), and the number that are simultaneously countable is again limited to roughly  $Z$  or  $Z^2/10$ . Events that are rare and sparsely represented cannot be efficiently counted, though if their probability rises, they may become countable in a different epoch (see section 6.4).

The number of simultaneously countable events is dramatically fewer than the  $Z^2$  distinct events that can be unambiguously represented on  $Z$  cells, and it scales with only the first or second power of  $Z$  not exponentially. This parallels long-established results on the storage and recall of patterns using synaptic weight changes, where the number of retrievable patterns scales between the first and second power of the number of cells, with the retrievable information (in bits) ranging from 8% to 69% (with extreme as-

sumptions) of the number of synapses involved for different models (e.g., Willshaw et al., 1969; Gardner-Medwin, 1976; Hopfield, 1982). Huge levels of redundancy are required in order simultaneously to count or learn about a large number ( $n$ ) of events, compared with the minimum number of cells ( $\log_2(n)$ ) on which these events could be distinctly represented. For  $Z = n$ , as required either for direct representations or for distributed representations with 50% counting efficiency on our projection model, the redundancy defined in this way is 93% for  $n = 100$ , rising to 99.9% for  $n = 10^4$ . This problem will reemerge in various guises in the following sections.

**6.2 Principles for Minimizing Counting Errors.** Our results lead to the following principles for achieving effective counting in distributed representations:

1. Where many events are to be counted on a net, use highly redundant codes with many times the number of cells needed simply to distinguish these events.
2. Use sparse coding for common and unimportant events, but raise the activity ratio for events that can be identified as likely to have useful associations, especially if the events are rare.
3. Minimize overlap between representations, aiming for overlaps less than those of the random selections of active cells that we have assumed for analysis.

These principles arise directly from the analysis of the problem as we have defined it: to count particular reproducible events on a network in the context of interfering events represented on the same network. Two more principles arise from considering how representations can economically permit counting that is effective for learning:

4. Representations should be organized so that counting can occur in small modules, each being a focus of information likely to be relevant to a particular type of association. Large numbers of events would then be grouped for counting into subsets (those that have the same representation in a module).
5. The special subsets that activate individual cells should, where possible, be ones that resemble each other in their associations. Overlap between representations should ideally mirror similarities in the implications of events in the external world.

These extra principles can help to avoid the necessity of counting large numbers of individual events, making learning depend on the counting of fewer categorizations of the environment, and therefore manageable on fewer cells. They can lead to appropriate generalization of learning and

can make learning faster since events within a subset obviously occur more often than the individual events.

**6.3 Does the Brain Follow These Principles?** One of the puzzles about the cortical representation of sensory information lies in the enormous number of cells that appear to be devoted to this purpose. A region of  $1 \text{ deg}^2$  at the human fovea contains about 10,000 sample points at the Nyquist interval (taking the spatial frequency limit as 50 cycles per degree), and the number of retinal cones sampling the image is quite close to this figure. But the number of cells in the primary visual cortex devoted to that region is of order  $10^8$ . Some of the  $10^4$ -fold increase may be explained by the role of the primary visual cortex in distributing information to many destinations, but this cannot account for all of it, and one must conclude that this cortical representation is grossly redundant. The selective advantage that has driven the evolution of cortical redundancy may be the necessity for efficient counting and learning, as encapsulated in our first principle.

In relation to the second principle, evidence for sparse coding was put forward by Legéndy and Salcman (1985), and it was clear from the earliest recordings from single neurons in the cerebral cortex that they are hard to excite vigorously and must spend most of their time firing at very low rates, with only brief periods of strong activity when they happen to be excited by their appropriate trigger feature. Field (1994), Baddeley (1996), Baddeley et al. (1997), Olshausen and Field (1997), van Hateren and van der Schaaf (1998), Smyth, Tolhurst, Baker, and Thompson (1998), Tolhurst, Smyth, Baker, and Thompson (1999), Tadmor and Tolhurst (in press), and others have now quantitatively substantiated this impression.

Counting accuracy for a particular event benefits from the sparse coding of other events (reducing their overlap and interference), not from its own sparse coding. Sparseness is especially important for common events because of the extent of interference they can cause, and because it can be afforded with little impact on their own counting, which is already quick and efficient. The benefits of sparseness are greatest for counting based on Hebbian synaptic modification (as in our support model and a related model for direct detection of associations: Gardner-Medwin & Barlow, 1992). Significant events (particularly rare ones) may need higher activity ratios to boost their counting efficiency at the expense of others. We envisage that this might be achieved if selective attention to important and novel events favors the representation of more features of the environment through lowering of cell thresholds and the acquisition of more detailed sensory information. Adaptation and habituation, on the other hand, can raise thresholds, favoring the required sparse representation of common events. This flexibility of distributed representations is attractive, for in biological circumstances, the probability and significance of individual events are highly context dependent.

Overlap reduction, as advocated in our third principle, follows to some extent from any mechanism achieving sparse coding. But there are also indications that well-known phenomena like the waterfall illusion and other after-effects of pattern adaptation involve a “repulsion” mechanism (Barlow, 1990) that would directly reduce overlap between representations following persistent coactivation of their elements (Barlow & Földiák, 1989; Földiák, 1990; Carandini, Barlow, O’Keefe, Poirson, & Movshon, 1997). Similar after-effects offer possible mechanisms for improving the long-term recall of overlapping representations of events, through processing during sleep (Gardner-Medwin, 1989).

The fourth principle fits rather well with the organization of the cortex in modules at several scales. The great majority of interactions between cortical cells are with other cells that are close by (Braitenberg & Schüz, 1991), so the smallest module might be a mini-column or column, such as are found in sensory areas (Edelman & Mountcastle, 1978). These would be the focus of the statistical structure comprising local correlations. Each pyramidal cell receives many (predominantly local) inputs that effectively define its module, as the connections to the accumulator cell define the network in our projection model (see Figure 1). The outputs then go to other areas of specialization, for example, area MT, as a focus for the spatiotemporal correlations of motion information. Optimal organization of a large system must involve mixing diverse forms of information to find new types of association (Barlow, 1981), as well as concentrating information that has revealed associations in the past, and this seems broadly consistent with cortical structure.

The trigger features of cortical neurons often make sense in terms of behaviorally important objects in the world surrounding an animal, suggesting that the brain exploits the advantages expressed in the fifth principle. For instance, complex cells in V1 respond to the same spatiotemporal pattern at several places in the receptive field, effectively generalizing for position. The same is true for motion selectivity in MT, and perhaps for cells that respond to faces or hands in inferotemporal cortex (Gross, Desimone, Albright, & Schwarz, 1985). Such direct representation of significant features (Barlow, 1972, 1995) assists in making distributed representations workable.

**6.4 The Relative Merits of Counting with Direct and Distributed Representations.** Maximum counting efficiencies for events represented within a network or counting module would be achieved with direct representations, but this requires prespecification of the patterns to be detected and counted. In contrast, a distributed network has the flexibility to represent and count, without modification, events that have not been foreseen. Our results show that provided these unforeseen representations have no more than chance levels of overlap with those of other events (as in our models), good performance can be achieved on a network of  $Z$  cells with almost any

limited selection (numbering of order  $Z$  or  $Z^2/10$ ) of frequent events out of the very large number (of order  $2^Z$ ) that can be represented. Infrequent but important events can be included within this selection by raising their activity ratios.

The potential for unambiguous representation of a huge number of rare and unforeseen events means that distributed networks can be very versatile for counting purposes in a nonstationary environment. Learning often takes place in short epochs during which some type of experience is frequent—for example, learning on a summer walk that nettles sting or enlarging one's vocabulary in a French lesson. A system in which there is gradual decay of interference from previous epochs, reasonably matched in its time course to the duration of a significant counting epoch, can in principle allow efficient counting of any transiently frequent event that is represented by any one of the patterns that the network can generate. This can assist in learning associations from short periods of intense and novel experience, though, of course, there may be hazards in generalizing such associations to other periods.

Once direct representations are set up for significant events, many such events might in principle be simultaneously detected and counted, through parallel processing. This cannot occur where events have overlapping distributed representations. One cannot have two different patterns simultaneously on the same set of cells. Thus, there is a trade-off between the flexibility of distributed representations in handling unforeseen events and the constraint that they can handle them only one at a time. Since events of importance at the sensory periphery are not generally mutually exclusive, it seems necessary to use distributed representations in a serial rather than a parallel fashion by attending to one aspect of the environment at a time. This is a significant cost to be paid for the versatility of distributed representations, but it resembles in some degree the way we handle novel and complex experience.

**6.5 Managing the Combinatorial Problem.** Huge numbers of high-level feature detectors are required by some models of sensory coding based on direct representations—the so-called grandmother cell or yellow volkswagen problem (Harris, 1980; Lettvin, 1995). The ability of distributed representations to represent a vastly greater number of events than is possible with direct representation is often thought to permit a way around this problem, but our results show that economy is not so simply achieved. For counting and learning, distributed representations have no advantage (or with our support model, rather limited advantage) over direct representations, because the maximum number that can be efficiently counted scales only with  $Z$  or  $Z^2$  rather than  $2^Z$ . Where distributed networks are used for representing larger numbers of significant events, counting on subsets of events (our fourth principle) may permit an economy of cells. This economy relies, however, on being able to separate off into a relatively small module the information that is needed to establish one type of association about an

event, while other modules receive information relevant to other associations. A combination of economical representation of events on distributed networks with the more extravagant use of cells required for counting may depend for success on a property of the represented events: that their associations can be learned and generalized from a separable fraction of the information they convey.

The combinatorial problem arises in an acute form at a sensory surface such as the retina, since impossibly large numbers of patterns can (and do) occur on small numbers of cells. Receptors that are not grouped close together experience states that are largely independent in a detailed scene, and for just 50 receptors, many of the  $10^{15}$  distinct events may be almost equally likely (even considering just two states per cell). This means that it is out of the question to count peripheral sensory events separately, and it would probably be impossible to identify even the correlations due to image structure without the topographic organization that allows this to be done initially on small sets of cells. The considerable (10,000-fold) expansion in cell numbers from retina to cortex (see section 6.3) allows for the counting of many subsets of events, but it would not go far toward efficient counting of useful events without using a hierarchy of small modules each analyzing different forms of structure to be found in the image.

## 7 Conclusion

---

Our analysis suggests ways in which distributed and direct representations should be related, and it has implications for understanding many features of the cortex. These include the expansion of cell numbers involved in cortical sensory representations, the extensive intracortical connectivity and its modular organization, the tendency for trigger features to correspond to behaviorally meaningful subsets of events, phenomena of habituation to common events, alerting to rare events, and attention to one thing at a time. Distributed representation is unavoidable in the brain but may cause serious errors in counting and inefficiency in learning unless it is guided by the principles that we have identified.

## References

---

- Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *Network*, 3, 213–251.
- Attneave, F. (1954). Informational aspects of visual perception. *Psychological Review*, 61, 183–193.
- Baddeley, R. J. (1996). An efficient code in V1? *Nature (London)*, 381, 560–561.
- Baddeley, R., Abbott, L. F., Booth, M. C. A., Sengpiel, F., Freeman, T., Wakeman, E. A., & Rolls, E. T. (1997). Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proceedings of the Royal Society, Series B*, 264, 1775–1783.

- Barlow, H. B. (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. In *The Mechanisation of thought processes* (pp. 535–559). London: Her Majesty's Stationery Office.
- Barlow, H. B. (1962). Measurements of the quantum efficiency of discrimination in human scotopic vision. *Journal of Physiology (London)*, *160*, 169–188.
- Barlow, H. B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, *1*, 371–394.
- Barlow, H. B. (1981). Critical limiting factors in the design of the eye and visual cortex. The Ferrier lecture, 1980. *Proceedings of the Royal Society, London, B*, *212*, 1–34.
- Barlow, H. B. (1989). Unsupervised learning. *Neural Computation*, *1*, 295–311.
- Barlow, H. B. (1990). A theory about the functional role and synaptic mechanism of visual after-effects. In C. B. Blakemore (Ed.), *Vision: Coding and efficiency*. Cambridge: Cambridge University Press.
- Barlow, H. B. (1995). The neuron doctrine in perception. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 415–435). Cambridge, MA: MIT Press.
- Barlow, H. B., & Földiák, P. (1989). Adaptation and decorrelation in the cortex. In R. Durbin, C. Miall, & G. Mitchison (Eds.), *The computing neuron* (pp. 54–72). Reading, MA: Addison-Wesley.
- Barlow, H. B., & Reeves, H. B. (1979). The versatility and absolute efficiency of detecting mirror symmetry in random dot displays. *Vision Research*, *19*, 783–793.
- Barlow, H. B., & Tripathy, S. P. (1997). Correspondence noise and signal pooling in the detection of coherent visual motion. *Journal of Neuroscience*, *17*, 7954–7966.
- Braitenberg, V., & Schüz, A. (1991). *Anatomy of the cortex: Statistics and geometry*. Berlin: Springer-Verlag.
- Carandini, M., Barlow, H. B., O'Keefe, L. P., Poirson, A. B., & Movshon, J. A. (1997). Adaptation to contingencies in macaque primary visual cortex. *Proceedings of the Royal Society, Series B*, *352*, 1149–1154.
- Churchland, P. S. (1986). *Neurophilosophy: Towards a unified science of the mind-brain*. Cambridge, MA: MIT Press.
- Edelman, G. E., & Mountcastle, V. B. (1978). *The mindful brain*. Cambridge, MA: MIT Press.
- Farah, M. J. (1994). Neuropsychological inference with an interactive brain: A critique of the "locality" assumption. *Behavioural and Brain Sciences*, *17*, 43–104.
- Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation*, *6*, 559–601.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.
- Földiák, P. (1990). Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics*, *64*(2), 165–170.
- Gardner-Medwin, A. R. (1969). Modifiable synapses necessary for learning. *Nature*, *223*, 916–919.
- Gardner-Medwin, A. R. (1976). The recall of events through the learning of associations between their parts. *Proceedings of the Royal Society B*, *194*, 375–402.

- Gardner-Medwin, A. R. (1989). Doubly modifiable synapses: A model of short and long-term auto-associative memory. *Proceedings of the Royal Society B*, *238*, 137–154.
- Gardner-Medwin, A. R., & Barlow, H. B. (1992). The effect of sparseness in distributed representations on the detectability of associations between sensory events. *Journal of Physiology*, *452*, 282P.
- Gross, C. G., Desimone, R., Albright, T. D., & Schwarz, E. L. (1985). Inferior temporal cortex and pattern recognition. In C. Chagas, R. Gattass, & C. Gross (Eds.), *Pattern recognition mechanisms* (pp. 165–178). Vatican City: Pontificia Academia Scientiarum.
- Harris, C. S. (1980). Insight or out of sight? Two examples of perceptual plasticity in the human adult. In C. S. Harris (Ed.), *Visual coding and adaptability* (pp. 95–149). Hillsdale, NJ: Erlbaum.
- Hinton, G. E., & Anderson, J. A. (1981). *Parallel models of associative memory*. Hillsdale, NJ: Erlbaum.
- Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, & the PDP Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 77–109). Cambridge, MA: MIT Press.
- Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational properties. *Proc. Natl. Acad. Sci. USA*, *79*, 2554–2558.
- Jones, R. C. (1959). Quantum efficiency of human vision. *Journal of the Optical Society of America*, *49*, 645–653.
- Legéndy, C. R., & Salzman, M. (1985). Bursts and recurrences of bursts in the spike trains of spontaneously active striate cortex neurons. *Journal of Neurophysiology*, *53*, 926–939.
- Lettvin, J. Y. (1995). J. Y. Lettvin on grandmother cells. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 434–435). Cambridge, MA: MIT Press.
- Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, *37*, 3311–3325.
- Page, M. (in press). Connectionist modeling in psychology: A localist manifesto. *Behavioural and Brain Sciences*, *23*.
- Rose, A. (1942). The relative sensitivities of television pick-up tubes, photographic film, and the human eye. *Proceedings of the Institute of Radio Engineers*, *30*, 293–300.
- Rumelhart, D. E., & McClelland, J. (1986). *Parallel distributed processing*. Cambridge, MA: MIT Press.
- Smyth, D., Tolhurst, D. J., Baker, G. E., & Thompson, I. D. (1998). Responses of neurons in the visual cortex of anaesthetized ferrets to natural visual scenes. *Journal of Physiology, London*, *509*, 50–51P.
- Tadmor, Y., & Tolhurst, D. J. (in press). The differences-of-gaussians' receptive field model and the contrasts in natural scenes. *Vision Research*.
- Tanner, W. P., & Birdsall, T. G. (1958). Definitions of  $d'$  and  $\eta$  as psychophysical measures. *Journal of the Acoustical Society of America*, *30*, 922–928.
- Tolhurst, D. J., Smyth, D., Baker, G. E., & Thompson, I. D. (1999). Variations in the sparseness of neuronal responses to natural scenes as recorded in striate cortex of anaesthetized ferrets. *Journal of Physiology, London*, *515*, 101P.

- van Hateren, F. H., & van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society, Series B*, 265, 359–366.
- Watanabe, S. (1960). Information-theoretical aspects of inductive and deductive inference. *I.B.M. Journal of Research and Development*, 4, 208–231.
- Willshaw, D. J., Buneman, O. P., & Longuet-Higgins, H. C. (1969). Non-holographic associative memory. *Nature*, 222, 960–962.

---

Received July 28, 1999; accepted June 14, 2000.