Available online at www.sciencedirect.com
ScienceDirect
Neural Networks 20 (2007) 564–576
Neural Networks
www.elsevier.com/locate/neunet

Optimal decision network with distributed representation

Rafal Bogacz*

Department of Computer Science, University of Bristol, Bristol BS8 1UB, UK

Received 22 October 2005; received in revised form 31 January 2007; accepted 31 January 2007

Abstract

On the basis of detailed analysis of reaction times and neurophysiological data from tasks involving choice, it has been proposed that the brain implements an optimal statistical test during simple perceptual decisions. It has been shown recently how this optimal test can be implemented in biologically plausible models of decision networks, but this analysis was restricted to very simplified localist models which include abstract units describing activity of whole cell assemblies rather than individual neurons. This paper derives the optimal parameters in a model of a decision network including individual neurons, in which the alternatives are represented by distributed patterns of neuronal activity. It is also shown how the optimal weights in the decision network can be learnt via iterative rules using information accessible for individual synapses. Simulations demonstrate that the network with the optimal synaptic weights achieves better performance and matches fundamental behavioural regularities observed in choice tasks (Hick's law and the relationship between the error rate and the time for decision) better than a network with synaptic weights set according to a standard Hebb rule.

Keywords: Decision making; Distributed representation; SPRT; Perceptual choice; Cell assemblies

1. Introduction

Experimental studies have shed light on the neural bases of simple perceptual decision-making and indicated that they involve at least three basic processes. First, sensory cortical areas provide noisy evidence in support of alternative choices (Britten, Shadlen, Newsome, & Movshon, 1993; Ditterich, Mazurek, & Shadlen, 2003; Hanks, Ditterich, & Shadlen, 2006). The noisy evidence supporting a particular alternative is represented in the firing rate of the sensory neurons selective for this alternative. Hence the goal of the decision process may be formulated as choosing the alternative for which the corresponding neuronal population has the highest mean firing rate (Gold & Shadlen, 2001, 2002). Second, it has been observed that in certain cortical regions (e.g. lateral intraparietal area (LIP) and frontal eye field (FEF)) neuronal firing rates gradually increase during the decision process, and it has been proposed that these areas integrate sensory evidence over time (Schall, 2001; Shadlen & Newsome, 2001). This integration averages out the noise present in the sensory evidence. Third, in the free-response paradigm in which animal can respond at any time, it has been observed that when the firing rate of these integrator neurons exceed a certain threshold, the decision is made and the action execution is initiated (Roitman & Shadlen, 2002).

The integration process during decision-making tasks takes a certain amount of time which is referred to as the *decision time* (DT). In tasks under the free-response paradigm, the reaction time consists of the DT and an additional period connected with visual and motor processes. DTs have been estimated from behavioural data (Ratcliff, Van Zandt, & McKoon, 1999; Usher & McClelland, 2001) and directly from neurophysiological data (Reddi, 2001; Sato, Murthy, Thompson, & Schall, 2001). In difficult tasks, the DT often constitutes the majority of the reaction time (e.g. Ratcliff et al. (1999)).

Due to the evolutionary pressure for the speed and the accuracy of choices, it may be plausible that the neural decision circuits operate in an optimal or nearly optimal way, i.e. minimizing the DT. Indeed, on the basis of careful studies of human DTs, psychologists (Laming, 1968; Ratcliff, 1978; Ratcliff et al., 1999; Stone, 1960) have proposed that during simple perceptual choice between two alternatives the brain effectively performs a sequential probability ratio test (SPRT) — an optimal algorithm allowing the fastest decisions for any required accuracy (Barnard, 1946; Wald, 1947; Wald

* Tel.: +44 117 954 5141; fax: +44 117 954 5208.
E-mail address: R.Bogacz@bristol.ac.uk.

0893-6080/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2007.01.003

Nomenclature

Throughout the paper the following notational convention is used: all variables in localist models are denoted by capital letters, while all variables in distributed models — by small letters.

A	number of alternative decisions
a	sparseness of coding in the distributed decision network
C	magnitude of noise in localist models
c	magnitude of noise in distributed stimuli
dt	integration constant
h_i^v	input to integrator neuron i via recurrent weights $v_{i,j}$
h_i^w	input to integrator neuron i via feedforward weights $w_{i,j}$
K	decay or leak in the localist decision network
k	decay or leak of integrator neurons
K_{UM}	decay or leak in the Usher and McClelland model
$l_{i,j}$	matrix of linear coefficients in the distributed decision network
M_I	mean input to unit I
n	number of neurons in each layer of the distributed decision network
V	weight of self-excitatory connections in the localist decision network
$v_{i,j}$	weights of connections between integrator neuron j and integrator neuron i
$w_{i,j}$	weights of connections between input neuron j and integrator neuron i
W_{INH}	weight of inhibitory connections in the localist decision network
$w_{\mathrm{inh},i}$	weights of connections between integrator and inhibitory neurons
W_{UM}	weight of inhibitory connections in the Usher and McClelland model
X_I	input to localist unit I
x_j	activity of input neuron j
$\mathbf{x}_{j,I}$	membership of input neuron j in assembly I
Y_I	activity level of localist unit I
y_i	activity of integrator neuron i
$\mathbf{y}_{i,I}$	membership of integrator neuron i in assembly I
α	learning rate
η_I	independent Wiener processes

& Wolfowitz, 1948). The theory postulating that the brain performs SPRT has also been shown to be consistent with neurophysiological data (Gold & Shadlen, 2001, 2002; Shadlen & Newsome, 2001; Smith & Ratcliff, 2004). This theory claims that the decision is made as soon as the *difference* between integrated evidence in support of the first and second alternative exceeds a positive or a negative threshold. It has been recently shown how SPRT may be implemented in biologically plausible neural network models in which the difference between the integrated evidence is computed via feedforward (Mazurek, Roitman, Ditterich, & Shadlen, 2003; Shadlen & Newsome, 2001) or feedback inhibitory connections (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Brown et al., 2005).

However, the above analyses were restricted to very simplified *localist models* which include abstract units or variables describing activity of whole cell assemblies rather than individual neurons. Cell assemblies (Hebb, 1949) are the distributed patterns of neuronal activity which represent visual stimuli in certain cortical areas (e.g. Gochin, Colombo, Dorfman, Gerstein, and Gross (1994). The computational models including separate units or variables describing activity of individual neurons are referred to as *distributed models*. Wang (2002) developed a biologically realistic distributed model of decision-making in area LIP (involved in controlling eye movements) during a task in which a monkey has to discriminate whether dots presented on the screen move left or right and to indicate its decision by making a saccade in the direction of movement. Recently, Wong and Wang (2006) have shown how SPRT may be implemented in the Wang (2002) model. In this model, each alternative decision is represented by a separate population of neurons, because in area LIP neurons selective for very different directions of saccade are located in separate locations of area LIP. Hence in the Wang (2002) model, the assemblies were non-overlapping: if a neuron belongs to the assembly representing one alternative, it does not belong to the assembly representing the other.

However, in the case of many other decisions, alternative choices are likely to be represented by overlapping assemblies so that a neuron selective for one alternative may also be selective for others. Many computational models have been proposed for how distributed overlapping representations may develop (e.g. Olshausen and Field (1996)) and be stored (e.g. Hopfield (1982)). Such overlapping distributed representations allow coding of many more alternatives by a group of neurons than non-overlapping representations. This property is particularly important in choices with many possible answers. For example, in the case of motor decisions, there is practically infinite number of possible movements, e.g. of an arm (including all combinations on angular velocities of all joints). And indeed, neurophysiological data suggest that the motor actions are encoded in distributed overlapping patterns of neuronal activity (e.g. Chapin (2004), Georgopoulos, DeLong, and Crutcher (1983), Georgopoulos, Pellizzer, Poliakov, and Schieber (1999), Schieber and Hibbard (1993)). Similar arguments apply to perceptual decisions (e.g. deciding what is the animal you are looking at, before choosing an appropriate action), which are the focus of this paper. During such perceptual decisions, alternative choices are also likely to be represented by overlapping assemblies, because complex visual stimuli are represented in this way in the visual areas in last stages of the ventral stream (Erickson, Jagadeesh, & Desimone, 2000) and in the prefrontal cortex (Averbeck, Crowe, Chafee, & Georgopoulos, 2003; Miller, Erickson, & Desimone, 1996).

The following experimental data suggest that during perceptual decisions: the sensory evidence is being integrated over time, and the alternatives are represented by overlapping patterns of neuronal activity. First, fMRI data indicate that prefrontal neurons integrate sensory information in the task in which human subjects were asked to decide whether a noisy stimulus is a face or a house (Heekeren, Marrett, Bandettini, & Underleider, 2004), and the prefrontal neurons have been shown to encode complex visual stimuli in overlapping distributed patterns of their activity (Averbeck et al., 2003; Miller et al., 1996). Second, DTs in tasks involving word discrimination have been well described by the models assuming evidence integration (e.g. Ratcliff (1978)), and the overlapping representations are likely to encode words as such representations encode vocalizations in primates (Romanski, Averbeck, & Diltz, 2005).

A response may not be required immediately after perceptual decisions (e.g. a predator may identify an animal and wait with the action) and the data show that the integration process is not limited to the tasks in which an immediate response is required: The gradually increasing firing rates have been observed in FEF also if there was a delay between stimulus offset and the response (Kim & Shadlen, 1999). Furthermore, in this study the FEF neurons represented the correct alternative even in the passive viewing condition when no response was required. Moreover, in the study of Gold and Shadlen (2003), the animals did not know which motor response was required for which alternative during stimulus viewing, and nevertheless, the accuracy increased with viewing time providing evidence for information integration.

Although the data reviewed above suggest plausibility of decision processes with overlapping distributed representations, there is neither experimental evidence demonstrating them directly nor theoretical work showing whether optimal decision making is possible with overlapping distributed representations. The latter theoretical question is addressed in this paper and the experiments that would confirm the existence of these decision processes are suggested in the discussion. This paper derives the values of the weights of connections in decision networks using distributed overlapping representation, which minimize DT for any fixed accuracy, and, in case of two alternatives, allow the network to implement SPRT. It is shown that these weights can be learnt via iterative rules using information accessible for individual synapses.

The derivation of the optimal parameters is achieved by finding the relationships between the parameters of the distributed and the localist decision networks for which both models perform exactly the same computations. Since the optimal parameters of the localist decision network are known, these relationships give the optimal parameters of the distributed decision network. But these relationships may also have value on their own as they bridge the two different levels of modelling of the neural circuits: localist and distributed. Hence they may be useful for deriving distributed versions also for other localist models, and thus grounding these models in the neurobiological implementation.

The model of decision network with distributed overlapping representation considered in this paper is not intended to map directly onto a particular cortical area, because the location of input neurons and integrator neurons may depend on the decision task. However in general, the input during this kind of perceptual decisions is likely to be provided by visual areas in the late ventral stream, representing complex features. The integration is likely to occur in frontal or in parietal areas, where it has been shown before (e.g. Heekeren et al. (2004), Shadlen and Newsome (2001)).

In this paper, the individual neurons are described as linear elements to enable mathematical tractability and to allow finding explicit conditions under which the localist and the distributed decision networks perform equivalent computations. The assumption of linearity in processing can be justified by assuming that attention acts to place non-linear integrators in the most sensitive, linear range of their response functions (e.g. Cohen, Dunbar, and McClelland (1990)).

The paper is organized as follows. Section 2 reviews localist models of decision making and conditions under which their performance is optimal. Section 3 derives an optimal distributed decision network. Section 4 shows that this network achieves faster decision times and matches behavioural data better than the networks in which the weights are set up according to a standard Hebb rule. Section 5 discusses the predictions of the theory and the direction of the future work.

2. Optimal localist decision networks

This section reviews three localist decision networks: the simplest one, i.e. the race model, and two optimal networks. Let A denote the number of alternative choices. All models reviewed here include A integrator units corresponding to assemblies accumulating evidence for each alternative. Let us denote the activity levels of these units (which may correspond to the total activity of neurons in cell assemblies) by Y_I ($I \in \{1, \ldots, A\}$). It is assumed that at the beginning of the decision process all $Y_I(0) = 0$. Each unit receives noisy input

$$X_I = M_I + C\eta_I \tag{1}$$

with mean M_I, and standard deviation C (η_I denote independent Wiener processes). All models reviewed here assume that whenever the activity of any unit exceeds a particular threshold, decision has been made in favour of the alternative represented by the first unit which crossed the threshold. The decision is considered to be correct if the mean input M_I to this unit is the highest among all units (this assumption is common in many models of decision making, e.g. Gurney, Prescot, and Redgrave (2001), Mazurek et al. (2003), Shadlen and Newsome (2001), Usher and McClelland (2001), Vickers (1970), Wang (2002)).

In the race model (Vickers, 1970) the units simply integrate their input X_I, so that the change in Y_I (or its derivative over time) is equal to: $\dot{Y}_I = X_I$. The race model, however, produces slower DTs than the model reviewed below (Bogacz et al., 2006; McMillen & Holmes, 2006), and is inconsistent with neurophysiological data: It predicts that integrators representing all alternatives should increase their activity during the integration process, while it has been observed that neurons representing the "losing" alternative decrease their activity (Shadlen & Newsome, 2001).

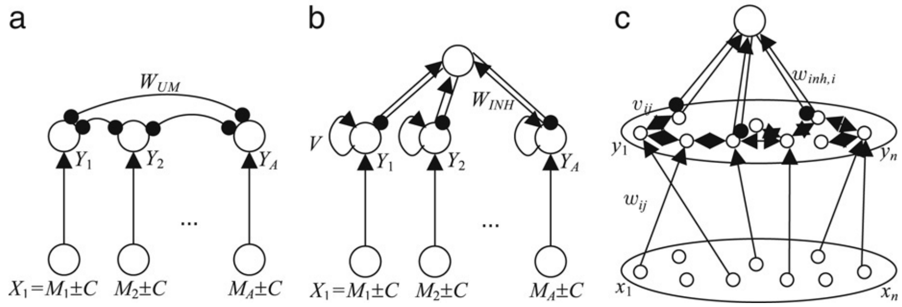

Fig. 1. The architectures of (a) the Usher and McClelland (2001), (b) the localist, and (c) the distributed decision networks. Small circles denote individual neurons, large circles denote cell assemblies. Arrows denote excitatory connections, lines ended with black circles denote inhibitory connections. For clarity, in panel (c) only sample connections are shown.

Usher and McClelland (2001) proposed a localist model of decision making with the architecture shown in Fig. 1(a). The changes in the activity of integrators are described by the following stochastic differential equations:

$$\dot{Y}_I = X_I - K_{\mathrm{UM}} Y_I - W_{\mathrm{UM}} \sum_{\substack{J=1 \\ J \neq I}}^{A} Y_J . \tag{2}$$

The activity levels of units decay (or leak) with proportionality constant K_{UM}. The model also includes competition between all units in the form of all-to-all inhibitory connections with weight W_{UM}.

Bogacz et al. (2006) have shown that for two alternatives ($A = 2$), when $K_{\mathrm{UM}} = W_{\mathrm{UM}}$ and the values of both these parameters are high (relative to input and noise), the performance of the Usher and McClelland (2001) model is optimal. For these parameters, the activity levels of the units are proportional to the difference between integrated evidence in support of the two alternatives, and hence the model approximates SPRT. As mentioned in the introduction, SPRT gives the fastest DTs for any required accuracy, which can be illustrated on the example of the race and the Usher and McClelland models. In both models, for given M_I and C, the DT and the accuracy depend on the height of the decision threshold. However, if the decision threshold is chosen in each of the models to give the same accuracy (e.g. 90%), then the optimally parameterized Usher and McClelland model will give faster average DT than the race model. Intuitively, this advantage of the Usher and McClelland model comes from its ability to adaptively react to the level of conflict between alternatives: If on a given trial the average input to the losing alternative is higher (due to noise), the winning unit will have to integrate for longer as it will receive more inhibition from the losing unit. Such adaptation of integration time is not present in the race model.

In case of multiple alternatives ($A > 2$), there exists a generalization of SPRT, known as Multihypothesis SPRT (Dragalin, Tertakovsky, & Veeravalli, 1999), but its implementation would require a network with architecture much more complex than that of the Usher and McClelland model. McMillen and Holmes (2006) have shown that for $A > 2$, the Usher and McClelland model achieves the lowest DT possible within this simple architecture also when $K_{\mathrm{UM}} = W_{\mathrm{UM}}$ and both these parameters are high. Let us call the parameters satisfying these constraints *optimal*, and in general, in this paper let us use the word *optimal* to refer to parameters that allow theoretically best possible performance (i.e. the shortest DTs for a fixed error rate) for $A = 2$, and that allow the best possible performance within the architecture considered for $A > 2$.

Wang (2002) proposed a detailed distributed model of decision making in area LIP. Fig. 1(b) shows a localist model with the architecture (connections between various neuronal populations) of the Wang (2002) model. It will be referred as the *localist decision network*. This model does not capture the complexity of the Wang (2002) model, but the simplifications made allow mathematical tractability. The localist decision network is very similar to the Usher and McClelland (2001) model with just two differences. First, the integrators do not inhibit one another, but rather send excitatory connections to a pool of inhibitory neurons, which then inhibit the integrators. Second, the integrator neurons send excitatory connections (denoted by V in Fig. 1(b)) within an assembly. These connections were found to be necessary to enable a network of individual model neurons, whose membrane voltages decay rapidly (on a millisecond scale), to integrate information on the timescales of decisions (hundreds of milliseconds). The changes in the activity of the localist units are described by the following stochastic differential equations (Bogacz et al., 2006):

$$\dot{Y}_I = X_I - K Y_I + V Y_I - W_{\mathrm{INH}} \sum_{J=1}^{A} Y_J . \tag{3}$$

To find the optimal parameters of the localist decision network, let us rewrite Eq. (3) as:

$$\dot{Y}_I = X_I - (K - V + W_{\mathrm{INH}})\, Y_I - W_{\mathrm{INH}} \sum_{\substack{J=1 \\ J \neq I}}^{A} Y_I . \tag{4}$$

Comparing Eqs. (2) and (4) shows that Usher and McClelland (2001) model and the localist decision network are computationally equivalent, when there are the following relationships between their parameters: $K_{\mathrm{UM}} = K - V + W_{\mathrm{INH}}$, $W_{\mathrm{UM}} = W_{\mathrm{INH}}$. Hence given that the optimal parameters of the Usher and McClelland model must satisfy: $K_{\mathrm{UM}} = W_{\mathrm{UM}}$ and both are high, the optimal parameters of the localist decision network must satisfy $K - V + W_{\mathrm{INH}} = W_{\mathrm{INH}}$, i.e. $K = V$, and W_{INH} is high (Bogacz et al., 2006).

3. Optimal distributed decision networks

This section derives a distributed decision network computationally equivalent to the optimal localist decision network. Let us define that two networks are *computationally equivalent* if for a given input both networks make the same choice after the same DT. This definition implies that the repeated simulation of two equivalent networks will yield exactly the same error rate and DT distribution. Since we know the optimal parameters of the localist decision network, the relationship between the parameters of the localist and the distributed decision networks giving computational equivalence will allow us to compute the optimal parameters of the distributed decision network.

3.1. Network architecture

The architecture of the distributed decision network is shown in Fig. 1(c). In this model, both layers of integrators and inputs are described as populations of n neurons. Let us denote the activities of the integrator neurons by y_i and the activities of the input neurons by x_j. Let us denote the weights of connections between input neuron j and integrator neuron i by $w_{i,j}$, and the weights of connections between integrator neuron j and integrator neuron i by $v_{i,j}$ (see Fig. 1(c)).

For simplicity, the inhibitory neurons are not modelled individually, but as a population, because they are not selective for different alternatives, by contrast to integrator and input neurons. Let us denote the weights of connections between integrator neuron i and the population of inhibitory neurons by $w_{\mathrm{inh},i}$, and for simplicity assume that the weights of connections between inhibitory and integrators neurons are equal to 1. Finally, let us denote the decay rate of the integrator neurons by k. For simplicity, let us model the individual integrator neurons as simple linear units. Hence the changes in the activity of these neurons are described by the following stochastic differential equations:

$$\dot{y}_i = \sum_{j=1}^{n} w_{i,j} x_j - k y_i + \sum_{j=1}^{n} v_{i,j} y_j - \sum_{j=1}^{n} w_{\mathrm{inh},j} y_j. \tag{5}$$

3.2. Distributed representation

Let us assume that each alternative I is represented by an assembly of an neurons (hence a denotes the sparseness of coding in the distributed network). Let us encode the relationship of integrator neurons belonging to assemblies in matrix $[\mathbf{y}_{i,I}]$, in particular: $\mathbf{y}_{i,I} = 1$ if neuron i belongs to assembly I, and $\mathbf{y}_{i,I} = 0$ otherwise (note that one neuron may belong to many assemblies). Similarly, let $\mathbf{x}_{j,I} = 1$ if neuron j belongs to assembly I, and $\mathbf{x}_{j,I} = 0$ otherwise. The theory described below works for any $0 < a < 1$, and also generalizes to a more biologically realistic case of neurons having different continuous responses to different stimuli (rather than responding or not as implied before). To represent this case the elements of matrices $[\mathbf{y}_{i,I}]$ and $[\mathbf{x}_{j,I}]$ would be continuous and would have to be normalized such that the average of each column is equal to a. However, for the clarity of argument, in the remainder of the paper only binary patterns will be used.

Let us now relate the variables of the localist and the distributed decision networks. It is most natural to assume that the activity of localist unit I corresponds to the total activity of neurons belonging to assembly I, which can be written as (all the variables indexed once, e.g. y_i, become column vectors):

$$[X_I] = \left[\mathbf{x}_{j,I}\right]^{\mathrm{T}} \left[x_j\right], \qquad [Y_I] = \left[\mathbf{y}_{i,I}\right]^{\mathrm{T}} [y_i]. \tag{6}$$

3.3. Parameters giving equivalence and optimal performance

To establish the equivalence between the localist and the distributed decision networks, we seek the relationships between parameters which will satisfy Eqs. (3), (5) and (6). Appendix A shows that if the following relationships are satisfied and pseudo-inverses of $[\mathbf{x}_{j,I}]^{\mathrm{T}}$ and $[\mathbf{y}_{i,I}]^{\mathrm{T}}$ exist, the computations of the localist and the distributed decision networks are equivalent:

$$k = K, \tag{7}$$

$$w_{\mathrm{inh},i} = W_{\mathrm{INH}} \frac{1}{an} \sum_{I=1}^{A} \mathbf{y}_{i,I}, \tag{8}$$

$$\left[v_{i,j}\right] = V \left[\mathbf{y}_{i,I}\right] \left[\mathbf{y}_{i,I}\right]^{-1}, \tag{9}$$

$$[w_{i,j}] = \left[\mathbf{y}_{i,I}\right]^{\mathrm{T}^{-1}} \left[\mathbf{x}_{i,I}\right]^{\mathrm{T}}. \tag{10}$$

Below the intuition is provided for why the above conditions need to be satisfied, and it is considered how the weights in the distributed decision network described by Eqs. (8)–(10) can be learnt in a biologically plausible manner, i.e. using only information accessible to individual synapses. It is usually assumed that a synapse (e.g. storing weight $v_{i,j}$) can only "access" information about activity of presynaptic and post synaptic neurons (e.g. y_i and y_j). Eqs. (9) and (10) seem to violate this condition, as they involve computation of pseudo-inverses (e.g. which requires an "access" to all elements of matrix $\mathbf{y}_{i,I}$), but it will be shown that there exist simple iterative learning algorithms using only information locally accessible to synapses that converge to the weights satisfying Eqs. (9) and (10).

The condition of Eq. (7) ensures that the individual neurons decay with the same rate as units in the localist model. Eq. (8) ensures that the inhibition received by each integrator neuron is equal to (left-hand side of this equation comes from Eq. (5), the two transformations use Eqs. (8) and (6)):

$$\sum_{j=1}^{n} w_{\mathrm{inh},j} y_j = \sum_{j=1}^{n} W_{\mathrm{INH}} \frac{1}{an} \sum_{I=1}^{A} \mathbf{y}_{j,I} y_j$$
$$= W_{\mathrm{INH}} \frac{1}{an} \sum_{I=1}^{A} Y_I. \tag{11}$$

Since there is *an* neurons in each assembly, the total inhibition received by all neurons in any assembly is *an* times higher, i.e. equal to $W_{\text{INH}} \sum_{J=1}^{A} Y_J$ yielding the equivalence with the localist model (compare with Eq. (3)). The weights of inhibitory connections of Eq. (8) could be learnt in the following way: they can be initialized to 0, and whenever a new assembly is introduced, the weights from the active integrator neurons to the inhibitory neurons are increased by W_{INH}/an.

Let us now consider the weights of connections between integrator neurons of Eq. (9). If one ignores constant V, Eq. (9) expresses the pseudo-inverse rule (Hertz, Krogh, & Palmer, 1991). To explore the properties of these weights, we first need to define the input to the integrator neuron i via the feedback weights as:

$$h_i^v = \sum_{j=1}^{n} v_{i,j} y_j. \tag{12}$$

The key property of the weights $v_{i,j}$ set according to the scaled pseudo-inverse rule is the following: if the integrator neurons are set to the pattern representing alternative I, i.e. $y_i = \mathbf{y}_{I,i}$, then the input h_i^v to each neuron i is equal to the activity of neuron i itself (Hertz et al., 1991) scaled by constant V:

$$h_i^v = V y_i. \tag{13}$$

Therefore, the feedback to the integrator neurons belonging to assembly I is equal to the activity of neurons in this assembly scaled by V, yielding equivalence with the localist model.

It has been shown that the weights $v_{i,j}$ can be learnt iteratively in the following procedure aiming to achieve the relationship expressed in Eq. (13) (Diederich & Opper, 1987; Hertz et al., 1991). One needs to repeat (until convergence) for every assembly I the following operations: (i) Set activities of integrator neurons to the pattern representing alternative I, and (ii) For every neuron i modify the weights in order to minimize the following cost function: Cost $= (-Vy_i + h_i^v)^2$. To minimize this cost function, the weights should be modified in the direction opposite to the gradient of Cost, i.e. Diederich and Opper (1987), Hertz et al. (1991):

$$\Delta v_{i,j} = \alpha \left(V y_i - h_i^v \right) y_j. \tag{14}$$

In Eq. (14) α denotes learning rate. Note the above equation uses only information locally available for an individual synapse, as the terms in the bracket correspond to a post-synaptic neuron, while y_j is the activity of a pre-synaptic neuron. Diederich and Opper (1987) showed that the above procedure converges to the weights satisfying Eq. (13), even if the integrator neurons do not project to themselves ($v_{i,i} = 0$).

Eq. (10) describing the weights between the inputs and the integrator neurons does not have any obvious biologically plausible implementation. However, it will be shown in the next subsection that making plausible assumption about patterns $\mathbf{y}_{i,I}$ simplifies Eq. (10) to the pseudo-inverse rule that, as explained above, can be easily learnt by a biological neural network.

When the parameters of the distributed decision networks are set according to Eqs. (7)–(10) and the underlying parameters of the localist network are set to optimal values ($V = K$, W_{INH} is high), then the performance of the distributed decision network is optimal for input patterns x_j satisfying Eqs. (1) and (6). In particular in this case, for $A = 2$, the distributed decision network implements the SPRT, and for $A > 2$, it achieves the best performance possible for the Usher and McClelland (2001) model.

3.4. *Weights of feedforward connections*

In this subsection we show that the rule for computing the feedforward weights of Eq. (10) simplifies to the biologically plausible pseudo-inverse rule, when it is assumed that for any two alternatives the similarity in their representations in the input layer is exactly the same (i.e. preserved) in their representations in the integration layer. To provide an intuition and the motivation for this assumption, we first explore the properties of the input to integrator neurons via the feedforward weights set according to Eq. (10). Then we prove the simplification of Eq. (10) to the pseudo-inverse rule, and finally, we discuss the possibility that the optimal feedforward weights can also be found by a large class of models of feature extraction.

3.4.1. *Dependence of input to integrator neurons on the overlap in representations*

Let us denote the input to the integrator neurons via feedforward weights, by:

$$h_i^w = \sum_{j=1}^{n} w_{i,j} x_j. \tag{15}$$

Fig. 2 explores the implications of Eq. (10) concerning the changes in h_i^w as a result of learning the representation of a new alternative. Each panel corresponds to one simulation and each simulation was performed as follows: initially, the input pattern $\mathbf{x}_{j,1}$ and the output pattern $\mathbf{y}_{i,1}$ corresponding to the first alternative were generated and are shown in rows 1 and 3 of the figure. Then, the feedforward weights were generated according to Eq. (10). Next, the input neurons were set to the first pattern $x_j = \mathbf{x}_{j,1}$, and the input h_i^w to the integrators was computed and it is shown in the fifth row of Fig. 2. Notice that h_i^w is equal to the representation of the first alternative in the integration layer, i.e. $h_i^w = \mathbf{y}_{i,1}$ (compare rows 3 and 5).

Then it was explored how this h_i^w changes after learning the representation of another alternative. Thus the input pattern $\mathbf{x}_{j,2}$ and the output pattern $\mathbf{y}_{i,2}$ corresponding to the second alternative were generated and are shown in rows 2 and 4 of Fig. 2. Then, the feedforward weights were generated according to Eq. (10), but this time using matrices $\mathbf{x}$ and $\mathbf{y}$ containing the representation of $A = 2$ alternatives. Next, the input neurons were set to the *first* pattern (as in the paragraph above) $x_j = \mathbf{x}_{j,1}$, and the input h_i^w to the integrator neurons was computed and it is shown in the sixth row of Fig. 2. As will be described in detail below, h_i^w sometimes changes and sometimes does not (compare rows 5 and 6) depending on the preservation of similarity between the input representations in the output representations.

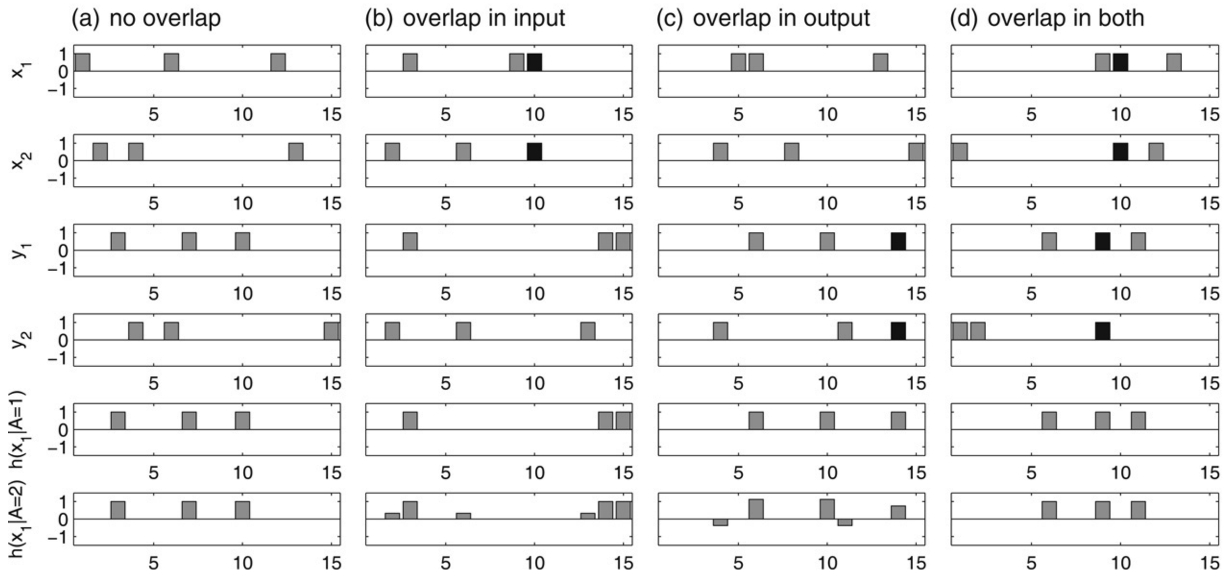

Fig. 2. The change in the input to integrator neurons as a result of the modification of feedforward weights. Each panel corresponds to one simulation and illustrates different effects (see text). In each simulation two input patterns and two output patterns of length $n = 15$ and sparseness $a = 0.2$ were generated randomly and are shown in four rows of the figure. The bottom two rows show inputs h_i^w to the integrators neurons after presentation of the first input pattern, when one pattern is stored in the weights (fifth row) and when two patterns are stored (sixth row; see main text for details).

In Fig. 2(a) there is no overlap between the representations of the two alternatives either in the input (compare rows 1 and 2) or in the output (compare rows 3 and 4). Since the patterns do not interfere with one another, it is not surprising that h_i^w also do not change after learning the representation of the second alternative.

By contrast in Fig. 2(b) the input representations of the two alternatives overlap in the 10th bit (overlapping bits are indicated by black bars), but there is no overlap in the output representations. Here, after learning, h_i^w increased its activity in the positions representing the second alternative (compare rows 4, 5, 6). This happens because $\mathbf{x}_{j,1}$ also provides evidence supporting the second alternative (if $x_j = \mathbf{x}_{j,1}$, then $X_1 = 4$, $X_2 = 1$ from Eq. (6)) and hence the integrator neurons representing the second alternative also receive the input.

Fig. 2(c) shows an opposite case in which there is no overlap in the input representations, but the output representations overlap in the 14th bit. In this case, h_i^w is *decreased* in the positions representing the second alternative (compare rows 4, 5, 6). This happens because $\mathbf{x}_{j,1}$ only provides the support for the first alternative ($X_1 = 4$, $X_2 = 0$), but unchanged h_i^w (as in row 5) would imply that the second alternative also receives input. Therefore, to counterbalance the overlap in the output representations, the integrator neurons representing the second alternative have decreased input.

The effects of overlap in the input and the integration layers are additive. In particular, when the numbers of overlapping bits in the input and the output representations are the same, the effects cancel each other, and h_i^w do not change during learning, as illustrated in Fig. 2(d). Let us now prove this property. The equality of the overlap in the input and the output representations for any two alternatives can be written as:

$$\left[\mathbf{x}_{j,I}\right]^{\mathrm{T}}\left[\mathbf{x}_{j,I}\right] = \left[\mathbf{y}_{i,I}\right]^{\mathrm{T}}\left[\mathbf{y}_{i,I}\right]. \tag{16}$$

The two sides of Eq. (16) contain matrices with the numbers of overlapping bits in all possible pairs of patterns. For, example in case of Fig. 2(d), these matrices are equal to $\begin{bmatrix}3 & 1\\ 1 & 3\end{bmatrix}$. Let us refer to the property described by Eq. (16) as *similarity preservation*.

We now show that if the similarity preservation is satisfied and the weights are computed from Eq. (10), then when input neurons encode pattern $\mathbf{x}_{j,I}$, the input to integrator neurons h_i^w is equal to the corresponding pattern $\mathbf{y}_{i,I}$ (in the calculation below, the first three transformations use Eqs. (15), (10) and (16) respectively),

$$\begin{aligned}\left[h_i^w\right] &= \left[w_{i,j}\right]\left[\mathbf{x}_{j,I}\right] = \left[\mathbf{y}_{i,I}\right]^{\mathrm{T}^{-1}}\left[\mathbf{x}_{j,I}\right]^{\mathrm{T}}\left[\mathbf{x}_{j,I}\right]\\ &= \left[\mathbf{y}_{i,I}\right]^{\mathrm{T}^{-1}}\left[\mathbf{y}_{i,I}\right]^{\mathrm{T}}\left[\mathbf{y}_{i,I}\right] = \left[\mathbf{y}_{i,I}\right].\end{aligned} \tag{17}$$

Let us notice that the property illustrated in Fig. 2(d) of h_i^w being preserved by learning the representation of a new alternative is a very desirable property. If it were not satisfied, then the response of the integrator neurons to the same input would change after learning of a new alternative, which would make the processing of the information from integrator neurons more difficult. Hence it would be beneficial for the neural system to choose representations satisfying the similarity preservation, and in fact we will show later that such representation is generated by a large class of feature extraction algorithms.

3.4.2. *Simplification of the optimal rule to the pseudo-inverse rule*

We now prove that when the similarity preservation is satisfied, then the rule for finding the feedforward weights of Eq. (10) simplifies to the pseudo-inverse rule (in the calculation below, the first transformation uses Eq. (10), and the third uses Eq. (16)).

$$[w_{i,j}] = \left[\mathbf{y}_{i,I}\right]^{\mathrm{T}^{-1}} \left[\mathbf{x}_{j,I}\right]^{\mathrm{T}} = \left[\mathbf{y}_{i,I}\right]^{\mathrm{T}^{-1}} \left[\mathbf{x}_{j,I}\right]^{\mathrm{T}} \left[\mathbf{x}_{j,I}\right] \left[\mathbf{x}_{j,I}\right]^{-1}$$
$$= \left[\mathbf{y}_{i,I}\right]^{\mathrm{T}-1} \left[\mathbf{y}_{i,I}\right]^{\mathrm{T}} \left[\mathbf{y}_{i,I}\right] \left[\mathbf{x}_{j,I}\right]^{-1}$$
$$= \left[\mathbf{y}_{i,I}\right] \left[\mathbf{x}_{j,I}\right]^{-1}. \tag{18}$$

The weights given by the right hand side of Eq. (18) can be computed by a biologically plausible iterative process using local information only (Hertz et al., 1991) analogous to that described in Section 3.3 for the feedback weights. Namely one needs to repeat (until convergence) for every assembly I the following operations: (i) Set the activities of integrator neurons to the pattern representing alternative I, and (ii) modify the feedforward weights according to:

$$\Delta w_{i,j} = \alpha \left(y_i - h_i^w\right) x_j. \tag{19}$$

3.4.3. *Finding optimal feedforward weights by feature extraction*

The iterative process described above could be used to find the weights satisfying Eq. (10) (i.e. the optimal weights) if the particular representation of alternatives satisfying the similarity preservation is somehow given by some external requirements. However, in the brain the representations in a certain processing stage often result from the process of feature extraction from the representations in the previous stage. Many feature extraction algorithms assume that the new representation can be computed by a linear feedforward network, and the algorithms learn the weights $w_{i,j}$ of the feedforward connections. In the context of the distributed decision network discussed here (in which the output neurons are the integrators), we could assume for simplicity that the output representation is equal to the input to the integrators via the feedforward weights (this would be the equilibrium state of the network given by Eq. (5) when neuronal decay $k = 1$, and other connections are not providing input):

$$\left[\mathbf{y}_{i,I}\right] = \left[w_{i,j}\right] \left[\mathbf{x}_{j,I}\right]. \tag{20}$$

In the previous part of this subsection we considered the case when we are given the representations of alternatives in both layers $\mathbf{x}_{j,I}$ and $\mathbf{y}_{i,I}$, and we sought weights $w_{i,j}$ associating them in an optimal way, while here let us consider the case in which we are given only the input representation $\mathbf{x}_{j,I}$, and we find weights $w_{i,j}$ via feature extraction, and $\mathbf{y}_{i,I}$ from Eq. (20).

Within the last decade it has been shown that models extracting features by a simple linear transformation of Eq. (20) may describe many properties of the features observed in the visual cortex (Bell & Sejnowski, 1997; Olshausen & Field, 1996). Neurophysiological data demonstrate that the representations of two similar stimuli are more similar than of two different stimuli (e.g. Freedman, Riesenhuber, Poggio, and Miller (2003), Kreiman, Koch, and Fried (2000)). Hence it is safe to assume that representations generated by feature extraction are likely to closely approximate the condition of similarity preservation. In fact, Appendix B shows that some more classical models of feature extraction give features satisfying similarity preservation exactly.

We now prove that the weights learnt by any linear feature extraction algorithm (i.e. producing features via Eq. (20)) generating the representation satisfying similarity preservation, also satisfy the condition required for optimal decision making of Eq. (10) (in the calculation below, the second transformation uses Eq. (20), and the third uses Eq. (18) which we proved to be correct when similarity preservation is satisfied):

$$\left[w_{i,j}\right] = \left[w_{i,j}\right] \left[\mathbf{x}_{j,I}\right] \left[\mathbf{x}_{j,I}\right]^{-1} = \left[\mathbf{y}_{i,I}\right] \left[\mathbf{x}_{j,I}\right]^{-1}$$
$$= \left[\mathbf{y}_{i,I}\right]^{T^{-1}} \left[\mathbf{x}_{j,I}\right]^{\mathrm{T}}. \tag{21}$$

This implies that the optimal feedforward weights for decision making of Eq. (10) can be closely approximated by biologically realistic feature extraction algorithms.

In summary, this subsection shows that the optimal values of the feedforward weights of the distributed decision network can be learnt using local learning rules in two ways: (i) if the distributed representations are "given" and they satisfy similarity preservation, the weights can be learnt using iterative version of the pseudo-inverse rule (Hertz et al., 1991); (ii) if the output representation can be chosen by the network itself, any feature extraction algorithm generating the representation satisfying similarity preservation will find the optimal weights.

4. Performance comparison

This section assesses the performance of the distributed decision network described in the previous section. First, Section 4.1 confirms in simulation that the distributed decision network achieves exactly the same error rates (ER) and DTs as the localist decision network with corresponding parameters and inputs. Then Sections 4.2 and 4.3 compare the performance and the match with behavioural data of the distributed decision network with the networks in which feedforward weights w or feedback weights v are set up according to a standard Hebb rule.

4.1. *Equivalence between localist and distributed networks*

To illustrate the equivalence between the localist and the distributed decision networks, the networks were simulated and their ER and DT were measured. The parameters of the localist network are given in the caption of Fig. 3, and the optimal parameters of the distributed network were derived as described in Section 3.3. In the simulated trials it was assumed that the first alternative is correct. Hence in each step of integration, the inputs x_j to the distributed network was equal to the first input pattern multiplied by a constant M_1 and integration step dt, with added Gaussian noise with variance c^2dt, i.e.

$$x_j = M_1 \mathbf{x}_{j,1} \mathrm{d}t + N(0, c^2 \mathrm{d}t), \tag{22}$$

where N denotes a random number sampled from a normal distribution with mean and variance given in brackets. Eq. (22) implies that the noise is independent between neurons. Although it is not the case in early sensory areas (e.g. Zohary,

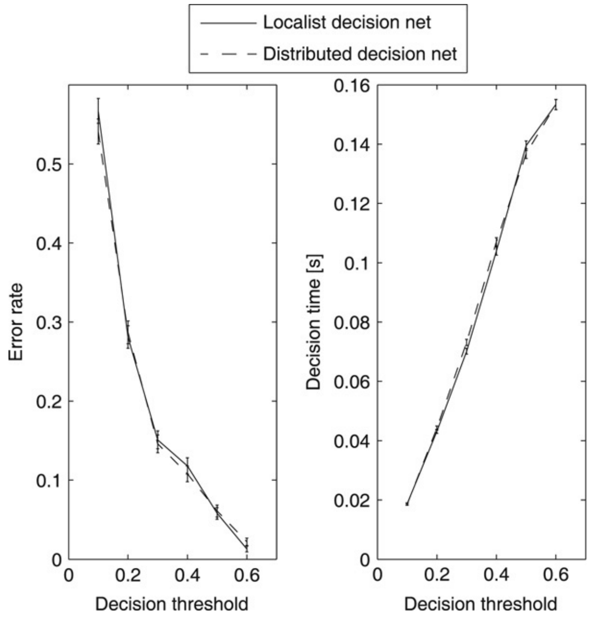

Fig. 3. Error rate and decision time (in seconds) of the localist decision network (solid line) and the distributed decision network (dashed line) for different values of decision threshold (shown on horizontal axes). The models were simulated using Euler method with dt = 0.01 [s]. The localist model was simulated with the following parameters: $A = 5$, $K = V = W_{\text{INH}} = 10$ [1/s]. The distributed model was simulated with $n = 20$, $a = 0.2$, and other parameters chosen to be equivalent to those of the localist model as described in Section 3.3. The weights of the integrator neurons were obtained from the iterative rules of Eqs. (14) and (19) (rather than 9, 10; it was assumes that $v_{i,i} = 0$). The inputs were generated as described in the main text with $M_1 = 1.41$ [1/s] and $c = 0.33$ [1/s] (the values estimated from sample participant of the experiment in study of Bogacz et al. (2006)). For each value of the decision threshold the simulations were repeated 1000 times and the error bars show the standard error.

Shadlen, and Newsome (1994)), in case of perceptual decisions considered in this paper, in which alternatives are represented by overlapping assemblies, the input is likely to be provided by the late areas in the ventral visual stream (see introduction). Erickson et al. (2000) report that in the last area in the ventral stream, the perirhinal cortex, the average correlation between noise in the responses of two neurons to the same stimuli is equal only to 0.017 (see Fig. 5(b) in Erickson et al. (2000)). So there, the noise processes are practically *un*correlated between neurons, as assumed in Eq. (22). The inputs to the localist decision network were generated on the basis of the inputs to the distributed network using Eq. (6).

Fig. 3 shows that neither ER nor DT differed significantly between the localist and the distributed decision networks. This result is not surprising, as the networks are shown to be equivalent in Section 3.3, but the result is given here to provide an independent validation of the equivalence.

4.2. Hebbian feedforward connections

We compare the performance of the distributed decision network with the network with feedforward weights $w_{i,j}$ set according to the Hebb rule for sparse associative memories (Amit, 1989) adapted for a hetero-associative memory (Kosko, 1988):

$$w_{i,j} = \frac{1}{na\,(1-a)} \sum_{I=1}^{A} \left(\mathbf{y}_{i,I} - a\right)\left(\mathbf{x}_{j,I} - a\right). \qquad (23)$$

The Hebb rule allows rapid learning based on a single presentation of a stimulus (the rules of the distributed network require iterative learning described in Section 3). The Hebb rule has the following property: if only one stimulus association is learnt by the network, i.e. $A = 1$, and it is presented on input $[x_j] = [\mathbf{x}_{j,1}]$, then the input to the integrator neurons is equal to the output representation of this stimulus decreased by a constant: $[h_i^w] = [\mathbf{y}_{i,1}] - a$ (analogous to the optimal rule; but for $A > 1$ this relation is only approximate (Amit, 1989)).

As the criterion of performance we choose DT for fixed ER (ER = 10%), i.e. the criterion optimized by SPRT (see Section 2). In the simulations, the underlying localist parameters (which occur in the equations for parameters of the distributed network) are set to satisfy the optimal values i.e. W_{inh} high and $K = V$ (see Section 2). To focus exclusively on the effect of the feedforward connections on performance (the effect of the feedback connections will be analysed in Section 4.3), here we assumed that the neurons are perfect integrators ($K = 0$) and hence the feedback connections are not present ($V = 0$). During the simulations, the network with Hebbian feedforward weights had inhibition set as described in Eq. (8). Fig. 4 shows that the network with Hebbian feedforward connections is slower than the optimal distributed decision network for a higher number of alternatives. Hence, it is useful for the brain to invest in more complex iterative learning (e.g. required by the optimal distributed decision network), as it reduces the DT.

The above result is not surprising, as the optimal distributed decision network by definition achieves better (or equal) performance than any other networks. However, the simulation shown in Fig. 4 allows the comparison of the network performance with behavioural data. In particular, the DT of the optimal distributed network is proportional to the logarithm of the number of alternatives A (note that there is a straight line in Fig. 4 with logarithmic scale for A), and thus follows Hick's law (Teichner & Krebs, 1974) widely observed in behavioural experiments involving choice between multiple alternatives. The optimal distributed network follows Hick's law, because it is equivalent to the localist decision network which has been shown to follow Hick's law (McMillen & Holmes, 2006). By contrast the network with Hebbian feedforward connections does not follow Hick's law (the points in Fig. 4 for the Hebbian network do not form a line).

4.3. Hebbian feedback connections

Let us now compare the optimal distributed decision network against the network with non-optimal ways of setting the feedback weights $v_{i,j}$ between the integrator neurons. First let us recall that the optimal way to set the feedback weights of Eq. (9) is the pseudo-inverse rule. We compare it against the

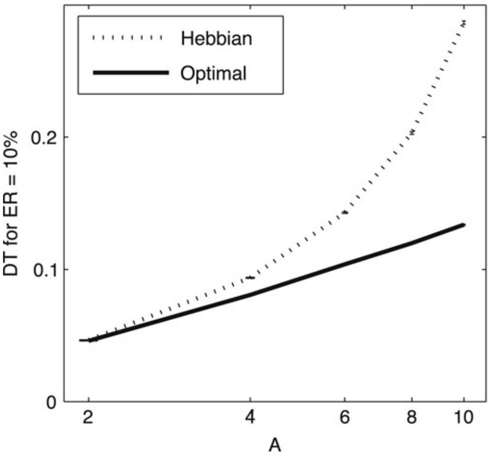

Fig. 4. Decision time (in seconds) of the optimal distributed decision network (solid line), and the network with feedforward Hebbian connections (dotted line) as a function of the number of alternatives A. The models were simulated using Euler method with $\mathrm{d}t = 0.01$ [s]. During all simulations the following parameters were kept constant: $K = V = 0$ [1/s], $W_{\mathrm{INH}} = 10$ [1/s], $n = 20$, $a = 0.2$, $M_1 = 1.41$ [1/s], $c = 0.33$ [1/s]. The number of alternatives A is shown on the horizontal axis. For each model and each number of alternatives, the decision threshold was found numerically that resulted in an error rate of 10% ± 0.2% (s.e.); this search for threshold was repeated 10 times. For each of these 10 thresholds, the decision time was then found in simulation and their average used to construct the data points. The standard error of mean decision time estimation was <2 ms for all data points; as indicated by very short error bars.

network with the feedback weights set according to the Hebb rule (Amit, 1989) analogous to that of Eq. (23):

$$v_{i,j} = \frac{V}{na\,(1-a)} \sum_{I=1}^{A} \left(\mathbf{y}_{i,I} - a\right)\left(\mathbf{y}_{j,I} - a\right). \tag{24}$$

To better illustrate the problems arising with the Hebb rule we use slightly different measure of model performance: ER for different times allowed for decision. This measure corresponds to the interrogation paradigm often used in psychological experiments (e.g. Usher and McClelland (2001)), in which participants are allowed a specified time for decision: from stimulus presentation to a special response cue presented at time T, after which the participants have to respond immediately. In the simulation of the interrogation protocol, the response at time T is considered to be correct if the most active assembly corresponds to the alternative with the highest mean input. Fig. 5 compares ER of the optimal distributed network, and the network in which the rule for setting feedback connections $v_{i,j}$ is changed to the Hebb rule of Eq. (24). The larger the values of decay parameter K and the strength of the feedback connections V, the more the performance of the network with Hebbian feedback weights departs from the optimal performance. To understand this result let us recall that the function of the feedback connections is to counterbalance the rapid decay of neuronal activity in the absence of input. If there is no such decay, the neurons integrate information perfectly, and there is no need for the feedback connections. The optimal rule of Eq. (9) is able to counterbalance the neuronal decay precisely, so the network integrates the input in the same way as the network without the neuronal decay and the feedback connections. The feedback weights set according to the Hebb rule do not counterbalance the neuronal decay as precisely (some assemblies have their level of activity maintained better than other). Hence the larger the strength of the feedback connections (necessary due to faster neuronal decay), the more imprecision is introduced to the integration process by the Hebbian feedback connections, and hence the larger is the ER.

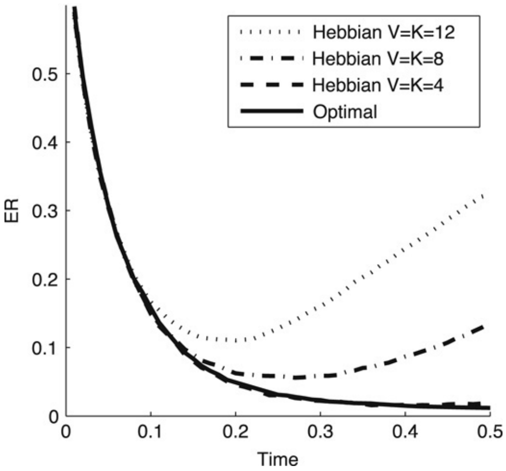

Fig. 5. Error rate in the interrogation paradigm of the optimal distributed network, and the network with Hebbian feedback connections, with different values of parameters V, K (shown in figure key in units of [1/s]). The models were simulated using Euler method with $\mathrm{d}t = 0.01$ [s]. During all simulations the following parameters were kept constant: $A = 5$, $W_{\mathrm{INH}} = 10$ [1/s], $M_1 = 1.41$ [1/s], $c = 0.33$, $n = 20$, $a = 0.2$. The time allowed for decision is shown on the horizontal axis in units of seconds.

Fig. 5 also shows that for the Hebbian feedback weights and for larger values of K and V (e.g. 8 and 12 in Fig. 5), the ER initially decreases but then starts to increase. The justification for this effect is given in Appendix C. The experimental results from the interrogation paradigm (Usher & McClelland, 2001) show that, as the time for decision increases, the ER of human participants decrease to a constant, as in the case of the optimal decision network in Fig. 5 (this constant may differ from 0 due to other factors not included in this simulation; see Usher and McClelland (2001)). The increase in ER with the increase of the time for decision, as in the case of the network with the Hebbian feedforward weights and with $V = K = 8$ or 12, is inconsistent with the experimental observations of Usher and McClelland (2001).

5. Discussion

This paper derives the optimal parameters of the decision network in which the alternatives are represented by neuronal cell assemblies, and shows that the optimal synaptic weights can be learnt by the rules using only information locally available to the synapses. Due to the evolutionary pressure for the speed and the accuracy of decisions, it is predicted that the parameters of biological decision networks will follow the derived optimal values. Simulations demonstrate that the distributed decision network with the optimal parameters is indeed consistent with Hick's law and the relationship between ER and the time for decision. The consistency with experimental observations is not present for the distributed decision network with the weights set up to the standard Hebb rule.

5.1. Experimental predictions

This paper has demonstrated that it is biologically plausible for the optimal decision networks with distributed representation to exist in brain, and this subsection describes experimental predictions that would confirm the existence of such networks.

The main reported manifestation of the decision processes is the gradual increase in the firing rate of the integrator neurons during presentation of noisy stimuli, thus similar increases should be observed in perceptual decisions based on noisy stimuli. These could be recorded from single neurons in certain frontal or parietal cortices during any task requiring discrimination between many complex visual stimuli (e.g. delayed matching to sample task (Miller et al., 1996), or recognition memory task (Xiang & Brown, 1998)). Initially, the stimuli (e.g. pictures of real world objects) could be presented without noise to establish the patterns of response of the neurons to individual stimuli. Then, the stimuli could be presented with different degrees of flickering white noise (as in a TV without an antenna) overlaid on top of them. The existence of decision networks with distributed representation predicts that there will exist integrator neurons with the following properties: (i) If the neurons are selective for a particular visual stimulus (when presented without noise), they will gradually increase their firing rate when this stimulus is presented with noise. (ii) The slope of this increase will be inversely proportional to the amount of the noise in the stimulus. (iii) The neurons (or at least some of them) should show these increases for more than one stimulus (as the distributed representations were assumed to be overlapping).

Furthermore, if the brain's distributed decision networks are equivalent to the Usher and McClelland (2001) model, as proposed in this paper, then these integrator neurons should show the dynamics similar to that of the localist units in the Usher and McClelland (2001) model. In particular, Usher and McClelland (2001) show that when two alternatives both receive high input, both integrator units initially increase, and then due to the mutual inhibition, the winner increases while the other decreases. Thus in the above mentioned experiment, if on some trials a noisy mixture of two stimuli is presented, then the firing rates of the neurons selective for the first stimulus and the neurons selective for the second should also initially increase, and then one population should increase and inhibit the other.

5.2. Linking with more detailed levels of description

This paper also describes iterative learning rules of Eqs. (14) and (19) resulting in the optimal weights, hence one can ask if they make any predictions regarding the synaptic plasticity. It is worth noticing that these rules do not fully specify the mechanism of plasticity, e.g. the postsynaptic neuron needs to provide the synapse with two values: y_i and h_i, and it is not obvious how it can be done, hence making the predictions about the synaptic plasticity may require a more precise model. One candidate for such a model could be the model of synaptic plasticity recently proposed by Norman, Newman, Detre, and Polyn (2006), since preliminary work of Xu (2006) suggests that it implements the learning described in Eqs. (14) and (19). In the Norman et al. (2006) model, the synapse can "access" $h_i - y_i$ due to the changes in the level of inhibition during theta oscillations.

In this paper only very simplified linear models of individual neurons were considered, and it would be interesting to extend the theory to more realistic description of neuronal behaviour. In particular, more experimental predictions would be generated by deriving the optimal parameters of a realistic distributed decision network with overlapping representations, as it has been done for non-overlapping representations (Wong & Wang, 2006). This paper is a first step in this direction.

Acknowledgments

This work is supported by EPSRC grant EP/C514416/1. The author thanks Roland Baddeley, Colin Davis and Marius Usher for discussion, and Jiaxiang Zhang, Andrew Lulham and Tobias Larsen for reading the previous version of the manuscript and very useful comments.

Appendix A. Optimal distributed decision network

This appendix shows that when the relationships of Eqs. (7)–(10) are satisfied, then the distributed decision network of Eq. (5) is equivalent to the localist decision network of Eq. (3).

Let us rewrite Eq. (3) in matrix notation. Let $[1]_{N,M}$ denote an N by M matrix filled with ones:

$$\left[\dot{Y}_I\right] = [X_I] - K\,[Y_I] + V\,[Y_I] - W_{\mathrm{INH}}\,[1]_{A,A}\,[Y_I]\,. \tag{25}$$

Let us rewrite Eq. (5) in matrix notation:

$$\begin{aligned}[\dot{y}_i] = {} & \left[w_{i,j}\right]\left[x_j\right] - k\,[y_i] + \left[v_{i,j}\right][y_i] \\ & - [1]_{n,1}\left[w_{\mathrm{inh},i}\right]^{\mathrm{T}}[y_i]\,. \end{aligned} \tag{26}$$

Let us rewrite Eq. (8) in matrix notation and transpose it:

$$\left[w_{\mathrm{inh},i}\right]^{\mathrm{T}} = W_{\mathrm{INH}}\frac{1}{an}\,[1]_{1,A}\left[\mathbf{y}_{i,I}\right]^{\mathrm{T}}. \tag{27}$$

Let us notice that matrix $[v_{i,j}]$ is symmetric (Hertz et al., 1991) so it is equal to its transpose, hence Eq. (9) can be rewritten as:

$$\left[v_{i,j}\right] = V\left[\mathbf{y}_{i,I}\right]^{\mathrm{T}^{-1}}\left[\mathbf{y}_{i,I}\right]^{\mathrm{T}}. \tag{28}$$

It will be now shown that when Eqs. (7), (9), (10), (27) and (28) are substituted into Eq. (26), then algebraic manipulations will yield Eq. (25).

Substituting Eqs. (7), (9), (10), (27) and (28) into Eq. (26), and using Eq. (6) we obtain:

$$\begin{aligned}[\dot{y}_i] = {} & \left[\mathbf{y}_{i,I}\right]^{\mathrm{T}^{-1}} [X_I] - K\,[y_i] + V\left[\mathbf{y}_{i,I}\right]^{\mathrm{T}^{-1}} [Y_I] \\ & - [1]_{n,1}\, W_{\mathrm{INH}} \frac{1}{an} [1]_{1,A}\, [Y_I]\,. \end{aligned} \tag{29}$$

Let us multiply the above equation by $[\mathbf{y}_{i,I}]^{\mathrm{T}}$. Notice that $[\mathbf{y}_{i,I}]^{\mathrm{T}}$ multiplied by vector $[1]_{n,1}$ (occurring in the last term of above equation) will result in a column vector of length A filled with an, because we assumed that each column of matrix $[\mathbf{y}_{i,I}]$ contains exactly an ones (each assembly consist of an neurons). We obtain:

$$\begin{aligned}\left[\dot{Y}_I\right] = {} & [X_I] - K\,[Y_I] + V\,[Y_I] \\ & - W_{\mathrm{INH}}\, [1]_{A,1}\, an \frac{1}{an} [1]_{1,A}\, [Y_I]\,. \end{aligned} \tag{30}$$

The multiplication and the cancellation in the last term of the above equation will result in Eq. (25).

Appendix B. Similarity preservation due to feature extraction

This appendix shows that the similarity preservation is satisfied by the representation generated by a set of classical models of feature extraction.

First we prove that the similarity preservation is satisfied when the columns of matrix $w_{i,j}$ with feedforward weights of the feature extraction network are orthonormal (in the calculation below, the first transformation uses Eq. (20), and the second transformation uses the orthonormality):

$$\begin{aligned}\left[\mathbf{y}_{i,I}\right]^{\mathrm{T}}\left[\mathbf{y}_{i,I}\right] &= \left[\mathbf{x}_{j,I}\right]^{\mathrm{T}}\left[w_{i,j}\right]^{\mathrm{T}}\left[w_{i,j}\right]\left[\mathbf{x}_{j,I}\right] \\ &= \left[\mathbf{x}_{j,I}\right]^{\mathrm{T}}\left[\mathbf{x}_{j,I}\right]. \end{aligned} \tag{31}$$

If the features were orthogonal but not normal, the Eq. (16) would be preserved modulo a scaling constant. The orthonormal weight $w_{i,j}$ are generated by the Principal Component Analysis which can be implemented by a number of models of feature extraction with local learning rules (e.g. Oja (1989), Sanger (1989)). Hence given the discussion in the end of Section 3.4, the weights of the feedforward connections generated by these rules would satisfy Eq. (10) precisely.

Appendix C. Dynamics with Hebbian feedback weights

This appendix provides justification for the effect observed in Fig. 5 that for the Hebbian feedback weights and for larger values of K and V (e.g. 8 and 12 in Fig. 5), the ER initially decreases but then starts to increase. To understand this result let us consider the linear coefficients of the Eq. (26) describing the dynamics of the distributed decision network; they are equal to $[l_{i,j}] = [v_{i,j}] - k[I] - [1]_{n,1}[w_{\mathrm{inh},i}]^{\mathrm{T}}$ (where $[I]$ denotes the identity matrix). Numerical explorations revealed that for the Hebbian feedback weights the matrix $[l_{i,j}]$ has some eigenvalues negative while some positive. Hence the fixed point of the deterministic part of Eq. (26) is a saddle point. Although the fixed point usually lies in the part of the state space corresponding to the correct alternative, the direction of the eigenvector corresponding to the highest eigenvalue does not depend on the input, thus it does not depend on which alternative is correct on a given trial. Hence initially, the system is attracted to the fixed point which is reflected by the decrease in ER, but then it is “captured” by the repulsion in the direction corresponding to the positive eigenvalues which is reflected by the increase in ER.

References

Amit, D. J. (1989). *Modelling brain function*. Cambridge: Cambridge University Press.

Averbeck, B. B., Crowe, D. A., Chafee, M. V., & Georgopoulos, A. P. (2003). Neural activity in prefrontal cortex during copying geometrical shapes. II. Decoding shape segments from neural ensembles. *Experimental Brain Research*, *150*, 142–153.

Barnard, G. A. (1946). Sequential tests in industrial statistics. *Journal of Royal Statistical Society Supplement*, *8*, 1–26.

Bell, A. J., & Sejnowski, T. J. (1997). The ”independent components” of natural scenes are edge filters. *Vision Research*, *37*, 3327–3338.

Bogacz, R., Brown, E. T., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of performance in two-alternative forced choice tasks. *Psychological Review*, *113*, 700–765.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1993). Responses of neurons in macaque MT to stochastic motion signals. *Visual Neuroscience*, *10*, 1157–1169.

Brown, E., Gao, J., Holmes, P., Bogacz, R., Gilzenrat, M., & Cohen, J. D. (2005). Simple networks that optimize decisions. *International Journal of Bifurcations and Chaos*, *15*, 803–826.

Chapin, J. K. (2004). Using muli-neuron population recordings for neural prosthetics. *Nature Neuroscience*, *7*, 452–455.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account for the Stroop effect. *Psychological Review*, *97*, 332–361.

Diederich, S. M., & Opper, M. (1987). Learning of correlated patterns in spin-glass networks by local learning rules. *Physical Review Letters*, *58*, 949–952.

Ditterich, J., Mazurek, M., & Shadlen, M. N. (2003). Microstimulation of visual cortex affects the speed of perceptual decisions. *Nature Neuroscience*, *6*, 891–898.

Dragalin, V. P., Tertakovsky, A. G., & Veeravalli, V. V. (1999). Multihypothesis sequential probability ratio tests — Part I: Asymptotic optimality. *IEEE Transactions on Information Theory*, *45*, 2448–2461.

Erickson, C. A., Jagadeesh, B., & Desimone, R. (2000). Clustering of perirhinal neurons with similar properties following visual experience in adult monkey. *Nature Neuroscience*, *3*, 1143–1148.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *Journal of Neuroscience*, *23*, 5235–5246.

Georgopoulos, A. P., DeLong, M. R., & Crutcher, M. D. (1983). Relations between parameters of step-tracking movements and single cell discharge in the globus pallidus and subthalamic nucleus of the behaving monkey. *Journal of Neuroscience*, *3*, 1586–1598.

Georgopoulos, A. P., Pellizzer, G., Poliakov, A. V., & Schieber, M. H. (1999). Neural coding of finger and wrist movement. *Journal of Computational Neuroscience*, *6*, 279–288.

Gochin, P. M., Colombo, M., Dorfman, G. A., Gerstein, G. L., & Gross, C. G. (1994). Neural ensemble coding in inferior temporal cortex. *Journal of Neurophysiology*, *71*, 2325–2337.

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, *5*, 10–16.

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions and reward. *Neuron*, *36*, 299–308.

Gold, J. I., & Shadlen, M. N. (2003). The influence of behavioral context on the representation of a perceptual decision in developing oculomotor commands. *Journal of Neuroscience*, *23*, 632–651.

Gurney, K., Prescot, T. J., & Redgrave, P. (2001). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics*, *84*, 401–410.

Hanks, T. D., Ditterich, J., & Shadlen, M. N. (2006). Microstimulation of macaque area LIP affects decision making in a motion discrimination task. *Nature Neuroscience*, *9*, 682–689.

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

Heekeren, H. R., Marrett, S., Bandettini, P. A., & Underleider, L. G. (2004). A general mechanism for perceptual decision making in the human brain. *Nature*, *431*, 859–862.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computations*. Redwood City, CA: Addison-Wesley.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceeding of the National Academy of Sciences*, *79*, 2554–2558.

Kim, J. -N., & Shadlen, M. N. (1999). Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nature Neuroscience*, *2*, 176–185.

Kosko, B. (1988). Bidirectional associative memories. *IEEE Transactions on Systems, Man and Cybernetics*, *18*, 49–60.

Kreiman, G., Koch, C., & Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience*, *3*, 946–953.

Laming, D. R. J. (1968). *Information theory of choice reaction time*. New York: Wiley.

Mazurek, M. E., Roitman, J. D., Ditterich, J., & Shadlen, M. N. (2003). A role for neural integrators in perceptual decision making. *Cerebral Cortex*, *13*, 1257–1269.

McMillen, T., & Holmes, P. (2006). The dynamics of choice among multiple alternatives. *Journal of Mathematical Psychology*, *50*, 30–57.

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience*, *16*, 5154–5167.

Norman, K. A., Newman, E., Detre, G., & Polyn, S. (2006). How inhibitory oscillations can train neural networks and punish competitors. *Neural Computation*, *18*, 1577–1610.

Oja, E. (1989). Neural networks, principal components and subspaces. *International Journal of Neural Systems*, *1*, 61–68.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607–609.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *83*, 59–108.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*, 261–300.

Reddi, B. A. J. (2001). Decision making: Two stages of neuronal judgement. *Current Biology*, *11*, R603–R606.

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience*, *22*, 9475–9489.

Romanski, L. M., Averbeck, B. B., & Diltz, M. (2005). Neural representation of vocalizations in the primate ventrolateral prefrontal cortex. *Journal of Neurophysiology*, *93*, 734–747.

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, *2*, 459–473.

Sato, T., Murthy, A., Thompson, K. G., & Schall, J. D. (2001). Search efficiency but not response interference affects visual selection in frontal eye field. *Neuron*, *30*, 1–20.

Schall, J. D. (2001). Neural basis of deciding, choosing and acting. *Nature Reviews Neuroscience*, *2*, 33–42.

Schieber, M. H., & Hibbard, L. S. (1993). How somatotopic is the motor cortex hand area? *Science*, *261*, 489–492.

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, *86*, 1916–1936.

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neuroscience*, *27*, 161–168.

Stone, M. (1960). Models for choice reaction time. *Psychometrika*, *25*, 251–260.

Teichner, W. H., & Krebs, M. J. (1974). Laws of visual choice reaction time. *Psychological Review*, *81*, 75–98.

Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review*, *108*, 550–592.

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics*, *13*, 37–59.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics*, *19*, 326–339.

Wang, X. -J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, *36*, 1–20.

Wong, K. -F., & Wang, X. -J. (2006). Exploring the neural mechanism of psychophysical reaction time: A dynamical analysis of a decision-making cortical microcircuit. *Journal of Neuroscience*, *26*, 1314–1328.

Xiang, J. Z., & Brown, M. W. (1998). Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. *Neuropharmacology*, *37*, 657–676.

Xu, W. (2006). Evaluating the model of familiarity discrimination by Norman et al. M.Sc. thesis, University of Bristol.

Zohary, E., Shadlen, M. N., & Newsome, W. T. (1994). Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, *370*, 140–143.