LETTER Communicated by Misha Tsodyks

The Hippocampus as a Stable Memory Allocator for Cortex

Leslie G. Valiant
valiant@seas.harvard.edu
School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, U.S.A.

It is suggested here that mammalian hippocampus serves as an allocator of neurons in cortex for memorizing new items. A construction of a shallow feedforward network with biologically plausible parameters is given that possesses the characteristics needed for such an allocator. In particular, the construction is stabilizing in that for inputs within a range of activity levels spanning more than an order of magnitude, the output will have activity levels differing as little as 1%. It is also noise tolerant in that pairs of input patterns that differ little will generate output patterns that differ little. Further, pairs of inputs that differ by much will be mapped to outputs that also differ sufficiently that they can be treated by cortex as distinct.

1 Introduction

The hippocampus is a part of the mammalian brain known to be essential for certain kinds of learning. It is widely believed that cortex rather than hippocampus is the main locus of information storage, but hippocampus is needed to help place certain kinds of memory into cortex. Experimental evidence from humans with hippocampal damage has been used to distinguish learning tasks for which the hippocampus is essential, from those for which it is not. Broadly speaking, the former has been characterized by terms such as *episodic* and *declarative learning* and the latter as *procedural learning* and *priming* (Scoville & Milner, 1957; O'Keefe & Nadel, 1978; Cohen, 1981; Tulving, 1983; Cohen & Eichenbaum, 1993; Squire, 1992; Schacter & Tulving, 1994; Schacter & Buckner, 1998; Martin, Schacter, Collins, & Rose, 2011).

Many attempts have been made to identify more explicitly the computational function of the hippocampus (e.g., Marr, 1971; Kali & Dayan, 2004; Teyler & DiScenna, 1986; Rolls, 1996). One important thread through some of these theories, apparently first articulated by Wickelgren (1979), is that the task that hippocampus enables is that of *chunking*, or the process of making a new concept out of a conjunction of concepts that are already stored separately, but not yet as a single concept in that combination. A second important thread, articulated by Teyler and DiScenna (1986), is that

the main function of hippocampus is to compute an "index" that facilitates the storage of information in cortex.

Here we consider the following integration of these two notions: the hippocampus enables conjunctive chunking in cortex by computing an index that facilitates this operation. Three questions need answering: What exactly is the index? How can it be computed with biologically plausible parameters? How does it facilitate chunking? The new hypothesis offered here is that a basic role of the hippocampus is to identify the set of neurons in cortex at which a new compound concept or chunk will be represented and to enable that set of neurons to take on that role. It is further suggested here that an important requirement is to ensure that the number of neurons allocated in this way to each new chunk is controlled within a limited range so as to avoid the overall system becoming unstable. For these reasons, we summarize this role of the hippocampus as that of a *stable memory allocator* (SMA). The technical content of this letter is the construction of circuits with biologically plausible parameters that possess the somewhat onerous properties that such an SMA needs.

The technical problem we solve is that of showing that circuits exist that will identify for any set T of active neurons within a wide range of sizes, say, where the largest is 10 times larger than the smallest, a set S of neurons, of close to, say within 1%, a fixed predetermined size. This will be achieved by a feedforward network of few, say three, layers where each layer is randomly connected. The first requirement of such a network is to stabilize the memory allocation process in this sense. However, there are two further requirements that we claim an SMA needs to have and that we show are also achieved by our construction. The first is that the system is noise tolerant in that a small enough fraction of neurons acting unreliably should not significantly degrade the working of the system. Our formulation of this requirement we call *continuity.* A complementary requirement we impose is that if a set S of neurons is assigned to a new item, then S should be substantially different from any S' that represents a different item if they are not to be confused. This property we call *orthogonality.*

We need all three properties to be realized by a single circuit that has numerical parameters with respect to neuron numbers, synapse numbers, synaptic strengths, activity level, or density of representation and the ratio of inhibition to excitation that are consistent with biology.

2 Chunking in the Neuroidal Model

If the main locus of memory is the much larger cortex and the hippocampus facilitates learning in cortex, no computational theory of hippocampus would appear to be meaningful that does not make reference to specific quantitative computational models of both cortex and hippocampus. To address this requirement, we adopt the neuroidal model (Valiant, 1994),

although our circuit constructions can be interpreted in other contexts too. While we accept Wickelgren's theory of what functionality hippocampus enables, namely chunking, we depart from his proposal for how that is being realized in cortex. He had proposed the cell assemblies of Hebb (1949), which apparently require that the neuron sets representing any one semantic item be particularly well interconnected. Explicit mechanisms for how such an assembly is co-opted when putting a new chunk into cortex have been lacking. In our formulation, no such special restrictions on the neuron sets are needed. The feasibility of all our claims can be demonstrated by explicit mechanisms in our model.

The neuroidal model offers the following interpretation of the chunking hypothesis of hippocampal function and the need for such a stabilizing mechanism. Suppose we want to identify the neuron set for a new conjunctive chunk $A\&B\& \ldots \&C$, when the constituents $A, B, \ldots, C$ are already represented by neuron sets in cortex. If the numbers of constituents of a chunk can vary, we clearly need some stabilizing mechanism. However, even if this number is fixed to, say, two, so that all chunks are of the form $A\&B$, we still need such a stabilizing mechanism if the number of neurons allocated to these chunks are not to vanish or explode when hierarchies of items of arbitrary depths of interdependencies are to be allocable.

For each *item* (e.g., an event, a concept), some set S of neurons is used to represent it. This representation comes in two varieties (Valiant, 2005). In the first, distinct items have *disjoint* sets of neurons. In the second, these sets can overlap so that individual neurons are *shared* among them.

The following semantics is needed for such a representation. If an item represented by S is being accessed during neural processing, then at least some fraction y of the S neurons must be firing, and if it is not being accessed, then less than some other fraction x must be firing. For example, in a particular large-scale simulation reported by Feldman and Valiant, 2009, the values $y = 88\%$ and $x = 30\%$ were used. This means that in the normal operation of the system, whenever more than 88% of S is firing, the item is definitely being accessed, and whenever less than 30%, it definitely is not. Most crucial, the system and algorithms are so configured that the situation when the fraction firing is in the intermediate range will occur extremely rarely. (A further detail in that simulation is that when an item is accessed, the possible fractions in that top 12% range occur with some probability distribution. For the negative examples, there is a similar distribution on the bottom 30%.)

As background to motivate the neuroidal model, we mention briefly here that a useful set of basic tasks has been shown to be supportable simultaneously (Valiant, 1994, 2005), and on the scale of tens and sometimes hundreds of thousands of task instances without substantial degradation (Feldman & Valiant, 2009). Furthermore, these demonstrations have been done on instances of this model designed to underestimate the numerical parameters of the brain and the capabilities of its components.

The basic tasks fall into two categories. First, there is a mechanism for allocating neurons to new items specified as the conjunction of items already represented. This is the operation that this letter addresses and corresponds to chunking. As described in the section that follows, stable mechanisms for this operation had been lacking. This is a gap that this letter fills.

Second, there is a basic set of tasks—association, supervised memorization of conjunctions, and inductive learning of certain linear separators—as defined by Valiant (1994) for which explicit algorithms have been given that perform them on items already represented in cortex. The purpose of such a suite of basic tasks is to enable complex knowledge structures to be built up in cortex. It is important that the basic tasks realized have in combination sufficient computational power to represent knowledge as required in cognition. This is to be contrasted with classical "associative memories" (Graham & Willshaw, 1997), which permit only flat structures and realize no more than the memorization and retrieval of bit sequences. The ability to represent new conjunctions of items, or chunks, as first-class objects adds fundamental power not available in flat structures such as classical associative memories. For example, it enables associations of conjunctions, such as $A\&B$, to a further item C, where no similar association to C from A or B separately is appropriate. The possibility of performing tasks, such as inductive learning or association, on features that are chunked items themselves of arbitrary depth of chunking provides expressive power not provided by single-task models such as classical associative memories.

3 The Stability Problem

In any theory of cortex where representations have some locality, one needs some explanation of how, for any item to be memorized, the actual set of neurons that represents it is determined. Previously suggested solutions to this problem have been mechanisms acting within cortex rather than in a separate device—the hippocampus—as we are suggesting here. Some have called the task of computing these addresses "recruitment learning" (Feldman, 1982; Diederich, Gunay, & Hogan, 2010). Valiant (1994) describes a particular mechanism called JOIN that provably performs the intended function of memory allocation. It assigns a set of neurons to a new item that is to be accessible as the conjunction of two previously memorized, say A and B. In particular, if the new item that represents $A\&B$ is C, then the neurons that are to represent C will be chosen to be the ones that are well connected to both A and B. This will permit the C nodes to be conveniently activated whenever both the A and B nodes are activated but not when just one of them is.

This JOIN mechanism is provably effective at any one level of activation (Valiant, 1994, 2005). Furthermore, simulation results show that one level of such activations is stable enough to support a variety of other tasks (Feldman & Valiant, 2009) on a large scale. However, as pointed out already

in Valiant (1994) and Gerbessiotis (2003), it has the weakness that when several allocation processes are performed in sequence, so that a deep hierarchy of allocations is created, the number of neurons allocated at successive depths will vary unstably, possibly ultimately vanishing to zero or filling up the cortex. In other words, if A and B have a desired number of neurons allocated, an allocation for $C = A\&B$ can be made. But as further levels are added, such as $D = C\&X$ and $E = D\&Y$, then control of the numbers allocated to D, E will be successively weaker and weaker.

The solution that Valiant (1994) suggested relied on the observation that among the tasks considered, memory allocation was the only one with a stability problem. Thus, if memory allocations are used only to some small enough depth, such as three, that the instability can be tolerated, then arbitrary data structures are still possible as long as these other links are the result of operations such as association that do not introduce instability. The suggestion made there was *naming*. If each item is allocated in the first instance according to the syntax or sound of a natural language word that describes it (rather than by its possibly arbitrarily deep semantic meaning), then allocating such words would require only a small fixed depth if the words consisted of at most a fixed number of syllables, say. (Implied here also was the psychological interpretation that naming a new concept is useful because it gives a ready way of allocating memory to it.) This letter suggests an analogous mechanism where now the names are not linguistic objects that need to be provided by the environment, but codes computed internally in the hippocampus, in analogy with hash functions in computing.

A different solution has been proposed by Gunay and Maida (2006; see also Diederich et al., 2010). They implement a stable version of JOIN within cortex that first allocates a number of neurons that exceeds the target size and then iteratively using inhibition reduces the number to the the target size. Our suggested solution appears much simpler and also achieves the further requirements that we set.

Beal and Knight (2008) suggest that instances of JOIN be realized at just one level and that by an association operation, these be linked to a separate item (created by an unspecified process) of controlled size.

The stability of neural networks in the firing frequency model has been widely investigated (Amari, 1974; Amit & Brunel, 1997; Tegne, Compte, & Wang, 2002; Latham & Nirenberg, 2004). However, these results do not appear to imply the existence of the circuits we seek, which generate a fixed output spiking pattern of stable size in a few steps and the further properties of continuity and orthogonality. Perhaps closer to our work is that of Minai and Levy (1994) and Smith, Wu, and Levy (2006) on recurrent networks of threshold elements that, like ours, are sparse, asymmetric, and motivated by the hippocampus. They show that stable activity levels can be achieved in principle, but doing so efficiently at the arbitrarily low levels we achieve here appears to be challenging in their setting.

4 A Theory of Hippocampus

Our suggestion is that a principal function of the hippocampus is to identify the set of neurons in cortex for new items that are conjunctions of previously stored items in a way that maintains stably the number of neurons allocated to every item. We proceed by first defining some properties that a device with this function would need to have and then show that there exist shallow feedforward circuits that possess these properties.

Our circuits have m input neurons and n output neurons, both large numbers such as 10^6, and for simplicity they are often equal. For input vector u of m 0/1 bits each, representing which input neurons are firing and which not in a certain time interval, we represent the outputs produced as vectors of n 0/1 bits, which specify which output neurons are firing. For vectors u, v, we represent their jth bits by u_j, v_j. The function computed by the device we call f, so that for input u, the output will be the n bit vector $f(u)$. The fraction of 1s in a vector u, the *density* of u, we denote by Dense(u), which measures the total activity level in a certain time interval. The number of bits on which two vectors u, v of the same length differ, or the Hamming distance, we denote by Ham(u, v). For the outputs Dense$(f(u))$ and Ham$(f(u), f(v))$ will denote the expected values of these quantities over randomization in the circuit construction. Also we denote the jth bits of u, v by u_j, v_j. We will represent by a the fraction of bits at which $u_j = v_j = 0$; b the fraction where $u_j = 0$, $v_j = 1$; c where $u_j = 1$, $v_j = 0$; and d where $u_j = 1, v_j = 1$. Hence $a + b + c + d = 1$.

We propose the following three properties as essential for an SMA:

1. *Stability.* For a wide range of input densities Dense(u), say, spanning the order-of-magnitude range from 0.002 to 0.025, we want the output density Dense$(f(u))$ to be in a narrow range, say, 0.0099 to 0.0101. We say that a circuit has $\epsilon - stability$ in the range $[q, s]$ if for some number p for any input u with density in that range, Dense$(f(u)) \in [p - \epsilon, p + \epsilon]$.
2. *Continuity.* If u and v are similar (i.e., Ham(u, v) small) enough that they should be regarded as close, noisy variants of each other, then $f(u)$ and $f(v)$ should also be similar enough (i.e., Ham$(f(u), f(v))$ small) that these outputs will be regarded by cortex as noisy variants of each other. For example, if u and v differ in a fraction of 10^{-4} neurons, then one might want $f(u)$ and $f(v)$ not to differ by more than, say, 10 times this quantity, or 10^{-3}. We say that a circuit has $\gamma - continuity$ in the range $[q, s]$ if for inputs u, v in that range of densities Ham$(f(u), f(v)) \leq \gamma$Ham(u, v).
3. *Orthogonality.* If u and v differ by enough (i.e., Ham(u, v) large) so that u and v should be regarded as distinct items, then the outputs $f(u)$ and $f(v)$ should differ sufficiently also (i.e., Ham$(f(u), f(v))$ large) that they will be regarded by cortex as distinct items. We say

that a circuit has $\delta - orthogonality$ in the range $[q, s]$ if for inputs u, v in that range of densities, $\mathrm{Ham}(f(u), f(v)) \geq \delta \mathrm{Ham}(u, v)$

The construction of our circuits uses randomization. Each circuit is determined by a vector $w \in W$ of real numbers that specifies its connections and weights and where w is drawn from W according to some probability distribution. Each resulting circuit will compute a (deterministic) function $f_w : \{0,1\}^m \to \{0,1\}^n$. Thus, the functions f_w can be regarded as the values of a random variable induced by the distribution on W, so that for some constant $\kappa > 0$, the properties that hold will be true with probability at least $1 - 2^{-\kappa n}$. For large enough n, the results will hold with overwhelming probability (i.e., the probability of failure is exponentially decreasing with n). We note that our results hold for every input pattern of firing—there are no inherently bad inputs. Each randomly constructed circuit will be bad for some extremely small fraction of inputs, the bad inputs depending on the connections that have been randomly chosen. As long as the random construction of the circuit is not correlated with the real-world experiences of the organism that determine the inputs, the bad inputs will occur so extremely rarely that they can be discounted for any practical purpose.

We note that using the above-defined notions, one can also analyze the rate at which the function f computed by the circuit changes as the circuit changes. In hippocampus, neurogenesis is believed to occur in adults in the dentate gyrus. If we regard this as the input layer of the circuit and regard neurogenesis as the replacement of neurons at a steady rate, then continuity and orthogonality are measures of the rate at which the function f changes with time.

5 Using the Stable Memory Allocator to Realize Memory Allocation in Cortex

Realizing memory allocation for a new conjunction $A\&B$ means assigning cortical neurons to the chunk $A\&B$ and changing cortical synapses as necessary so that in the future, whenever the cortical neurons that represent the constituent items A and B fire, so will the neurons assigned to that chunk. Furthermore, the chunk will be caused to fire by the firing of the sets A and B by means of a circuit entirely in cortex, with no further help needed from hippocampus.

The SMA can be used to realize the allocation in the following way. The task of *supervised memorization of conjunctions* was defined in Valiant (1994). For that task, we have neuron sets A, B, and C and can cause any of these three sets to fire at will for the purposes of a training process. The goal of the task is to set up a circuit such that in the future, when A and B fire, so will the neurons C. This is easier than the memory allocation problem addressed here in the two senses that neurons are already identified for C before the start and it is known how the set C can be caused to fire.

The point is that an SMA can be used to fill this gap in the following way. Suppose that the set of all the neurons in cortex is regarded as the input layer to the SMA and also as the output layer. Then the firing of neuron sets A and B in cortex, and hence also in the input layer of the SMA, will cause a (stable) set of neurons, which we shall call D, to fire in the output layer of the SMA, and hence also in cortex. This means that the SMA has identified a set D, and also that it gives a way of causing D to be fired at will via the SMA during any training process. This is equivalent to saying that memory allocation for $A\&B$ can be realized by first having the SMA identify the neuron set D and then training a circuit located entirely in cortex, with inputs A and B and output D to realize the supervised memorization of $A\&B$ at the node set D. The effect will be memory allocation for the chunk $A\&B$ at the node set D via a circuit entirely in cortex.

Another way of saying this is that the SMA enables the memory allocation problem to be reduced to the more easily supervised memorization problem. Now in Valiant (1994), it was shown that under the strong synapse hypothesis (single presynaptic neurons being able to cause a postsynaptic neuron to fire), supervised memorization can indeed be carried out in general for conjunctions of r items (where, furthermore, the algorithm does not need to know r) for $1 \leq r \leq 5$, for reasonable values of the other parameters. In the weak synapse regime, algorithms for supervised memorization have been given for this task for $r = 2$ (Valiant, 2005; Feldman & Valiant, 2009).

We can view the SMA as consisting of three networks S1, S2, and S3 connected in that order end to end where S1 is a down transformer with, say, 10^{10} inputs, and 10^7 outputs; S2 is the inner SMA with 10^7 inputs and 10^7 outputs; and S3 is the up transformer from 10^7 to 10^{10} neurons. For expository reasons in this letter, we are analyzing the inner SMA and simply comment that single-layer random circuits are enough to realize both the down transformation and the up transformation. However, the analysis we give can be easily combined to apply to the overall three-stage circuit having varying numbers of neurons at each stage.

The reason that the transformers are easy to implement is that all that is required of them is that they be low-variance circuits in the sense that a fixed density of activation of the inputs should give a fixed expected density of activation of the outputs, with low variance. (Note that the stability requirement is much more onerous since it makes demands on the output densities, not just one input density at a time.) Consider a circuit with m inputs and n outputs, with each output connected to k randomly chosen inputs. Suppose that each output is a threshold function that fires iff at least h of those k inputs are firing. Then if a fraction p of the inputs are firing, the probability that any one output fires will be $F(p)$ where F depends on k and h. If the connections to the different outputs are independent of each other, the number of outputs that fire will be governed by a binomial distribution with expectation $nF(p)$ and variance $nF(p)(1 - F(p))$. If we consider such a circuit to be an up transformer, from, say, 10^7 to, say, 10^{10} neurons, then

the expectation $nF(p)$, which corresponds to the number of neurons to be assigned to a concept in cortex, may be 50 or 10^5 (dependent on k and h), but the number of neurons that will be allocated will most likely differ from whatever the desired number is by no more than about the square root of that quantity, because of the low variance of the binomial distribution. In this sense, therefore, any randomly connected one-level circuit of threshold functions is a low-variance circuit. The same argument holds for down transformers.

While this letter advocates the position that a stabilizer for cortex is essential, we do not know the level of accuracy at which stabilizing is needed or which exactly are the neural layers that achieve it. For that reason, we refrain from identifying the neural correlates of the various parts of our construction in detail. However, the approximate intention is that the entorhinal cortex is the correlate of our down and up transformers, and the hippocampus the correlate of the inner SMA that we analyze in the sections to follow. But we do not make conjectures about which individual layers of hippocampus are contributing how much to stabilizing.

6 A Basic Construction and Its Analysis

We first consider a bipartite network with m inputs and $n = m$ outputs, where each output is connected to four inputs chosen independently at random (allowing repetitions). Consider in particular that each output realizes the threshold function $x + y + z - 2t \geq 1$ on the inputs to which it is connected. If fraction p of the inputs have value 1 rather than 0, then the probability that the output will have value 1 will be $h(p) = (1-p)(1-(1-p)^3) + p^4$. The first term represents the case that the t input is 0 so that any values of the other three variables will suffice except when all three have value zero. The second represents the case that the t input has value 1, which requires that each of the other three inputs also has value 1. We note that this expression equals

$$h(p) = 4p^3 - 6p^2 + 3p.$$

Suppose that one stacks these networks in sequence so that the input to the ith is the output to the $(i-1)$st and that we denote the fraction of outputs to the $i-1$st layer that have value 1 by p_{i-1}. Then the probability that any one bit of the output of the ith such network will have value 1 is $p = h(p_{i-1}) = 4p_{i-1}^3 - 6p_{i-1}^2 + 3p_{i-1}$. Suppose also that p^* is a fixed point of h, so that $p^* = h(p^*)$. Then for convergence of p under iteration of this bipartite construction, we need the magnitude of the derivative, $\alpha = |h'(p^*)|$, to be less than 1. That is a sufficient condition to ensure that for some interval $[q, s]$ with $q < p^* < s$ and for some constant $\beta < 1$, the iteration $p_i = h(p_{i-1})$ started from any p_0 in $[q, s]$, will converge toward p^* in the sense that

$|p^* - p_i| < \beta^i |p^* - p_0|$ for all i. (The reason is that $h(p) - h(p^*) = h'(p^*)(p - p^*) + o(p - p^*)$, and hence for any β with $\alpha < \beta < 1$, $|h(p) - h(p^*)| < \beta |p - p^*|$ for all sufficiently small $|p - p^*|$.) This suffices to establish that for every $\epsilon > 0$, there is an i such that i of these circuits stacked together is ϵ-stable in some interval.

In other words the conditions for stability that need to be verified are that (1) the equation $p = h(p) = 4p^3 - 6p^2 + 3p$ has a fixed point $p^* \in (0, 1)$, and (2) at that fixed point p, $|h'(p)| = |12p^2 - 12p + 3| < 1$. For condition 1, we need that $4p^3 - 6p^2 + 2p = 0$ have a solution in $(0, 1)$. In fact $p = 1/2$ is such a solution. For condition 2, we note that the derivative $h'(p) = 12p^2 - 12p + 3$ at $p = 1/2$ has value 0, which, being less than 1 in magnitude, guarantees convergence. It can be shown that for any $\epsilon > 0$ and for the range $[q, s]$ for any $0 < q < 0.5 < s < 1$, there is an i such that this circuit with i layers will have $\epsilon - stability$.

The reader can verify that many other threshold functions do not have such a convergence property. For example, $x + y \geq 1$, $x + y + z \geq 1$, $x + y + z + t \geq 2$ and $x + y + z - t \geq 2$ do not have fixed points in $(0, 1)$. Some others, such as $x + y + z \geq 2$ and $x + y + z + t \geq 2$, have such fixed points, but their derivatives $h'(p)$ there are larger than 1 in magnitude. A third category satisfies these two conditions but fails to satisfy the orthogonality condition. Such a case is $x + y - 2t \geq 1$, which, for the two very different inputs $u = 0^m$ and $v = 1^m$, will produce the same output 0^m.

However, a second effective threshold function is $x + y - t \geq 1$, and we will analyze this also. It turns out that this has worse convergence but better continuity than does $x + y + z - 2t \geq 1$. The analysis of convergence is similar to that given above and goes as follows. Now, $h(p) = (1 - p)(1 - (1 - p)^2) + p^3$. The first term represents the case that the t input is 0 so that any values of the other two variables will suffice except when both have value zero. The second represents the case that the t input has value 1, which requires that both of the other inputs also have value 1. This expression equals

$$h(p) = 2p^3 - 3p^2 + 2p.$$

The fixed point of the equation $h(p) = p$ in $(0, 1)$ is again $p^* = 1/2$, but at that point, the derivative $h'(p) = 6p^2 - 6p + 2$ equals $1/2$, which is sufficient for convergence, but at a slower rate than the zero derivative for the previous equation.

7 Analysis of Continuity and Orthogonality

We shall prove analytically that our construction based on $x + y + z - 2t \geq 1$ for depth i achieves 3^i-continuity in the interval $(0, 1)$. We shall also show

that $(3/2)^i$-orthogonality is achieved asymptotically as $\mathrm{Ham}(u, v) \to 0$ and $(0.719..)^i$-orthogonality throughout the range.

We first prove an upper bound of 3 on the continuity of each layer of the iterated $x + y + z - 2t \geq 1$ circuit. The argument will also serve as the basis for the analysis of orthogonality here, as well as for our analyses of constructions that have equilibria at the lower densities we consider later.

In any fixed position j in the input vectors u, v, the jth bit pair u_j, v_j has one of four combinations of values: 00,01,10,11. We have defined the fractions that these four possible combinations each accounts for among the m different values of j to be a, b, c, d, respectively, so that $a + b + c + d = 1$. Now each one of the four input connections to a particular output i can come from any one of these four regions, so there are $4^4 = 256$ possible combinations. For each of these 256 possibilities for the four connections to output i, we let $U = 1$ iff $x + y + z - 2t \geq 1$ hold for u and let $V = 1$ iff $x + y + z - 2t \geq 1$ hold for v. Now for any such fixed four connections to j and any fixed u, v, either $U = V$ or $U \neq V$. We now evaluate the total probability X of all the possibilities among the 256 that give the latter, that is, $U \neq V$, for the four connections to output i chosen randomly and u and v chosen arbitrarily with the fixed fractions a, b, c, and d.

For example, if the x, y, z connections all come from the 01 region and the t connection from the 00 region, then $U = 0$ while $V = 1$. Also the probability of this is ab^3. Hence, ab^3 will be a contribution to X.

Using similar arguments, partitioning the 256 cases according to whether the t connection comes from the 00, 01, 10, or 11 region, and further whether $U = 1$ & $V = 0$ or $U = 0$ & $V = 1$, the contributions to X in the eight cases can be verified to be:

$t = 0$ on u, $t = 0$ on v, $U = 0$, $V = 1 : a(b^3 + 3b^2a + 3ba^2)$.
$t = 0$ on u, $t = 0$ on v, $U = 1$, $V = 0 : a(c^3 + 3c^2a + 3ca^2)$.
$t = 1$ on u, $t = 1$ on v, $U = 0$, $V = 1 : d(b^3 + 3b^2d + 3bd^2)$.
$t = 1$ on u, $t = 1$ on v, $U = 1$, $V = 0 : d(c^3 + 3c^2d + 3cd^2)$.
$t = 0$ on u, $t = 1$ on v, $U = 0$, $V = 1 : b^4$.
$t = 0$ on u, $t = 1$ on v, $U = 1$, $V = 0 :$
$\quad b(1 - (a^3 + 3a^2b + 3b^2a + b^3 + 3b^2d + 3bd^2 + d^3))$.
$t = 1$ on u, $t = 0$ on v, $U = 0$, $V = 1 :$
$\quad c(1 - (a^3 + 3a^2c + 3c^2a + c^3 + 3c^2d + 3cd^2 + d^3))$.
$t = 1$ on u, $t = 0$ on v, $U = 1$, $V = 0 : c^4$.

In the sixth and seventh of these cases, it is easier to compute the probability that the claimed condition on x, y, z is violated and subtract it from 1. After some cancellation, the sum of these eight expressions evaluates to

$$X = (b + c)(1 + 2a^3 + 2d^3) - 2(a + d)(c^3 + b^3).$$

On substituting for the input disagreement $D = (b + c)$ and simplifying using $(a + b + c + d) = 1$, we get that the output disagreement is

$$X = D(1 + 2(a^3 + d^3 + D^3 - D^2 + 3bc(1 - D))).$$

We define the *expansion* E to the multiplicative increase in disagreement caused by the circuit, namely, X/D. Using $a + b + c + d = 1$, we deduce that $E = 1 + 2(a^3 + d^3 + D^3 - D^2 + 3bc(1 - D))$. Since for any D, $a^3 + d^3$ maximizes at $(1 - D)^3$ when one of a or d equals 0 and bc maximizes at $D^2/4$ when $b = c = D/2$, it follows that E is upper-bounded by $1 + 2((1 - D)^3 + D^3 - D^2 + 3D^2(1 - D)/4) = 3 - 6D + 11D^2/2 - 3D^3/2$, which is upper-bounded by 3 for all values of $D \in [0, 1]$. Hence for $x + y + z - 2t \geq 1$,

$$E \leq 3. \tag{7.1}$$

As far as the minimum expansion, since for any fixed D, $a^3 + d^3$ minimizes at $(1 - D)^3/4$ when $a = d = (1 - D)/2$, and bc minimizes at 0 when one of b or c is 0, it follows that $E = (1 + 2(a^3 + d^3 + D^3 - D^2 + 3bc(1 - D))) \geq 1 + (1 - D)^3/2 + 2D^3 - 2D^2 = (3 - 3D - D^2 + 3D^3)/2$. This worst-case situation, with one of b or c equal to zero, has expansion $E > 1$ for $D \in (0, 1/3)$, expansion $E = 1$ at $1/3$ and 1, and $0.719 < E < 1$ for $D \in (1/3, 1)$. We note that when D is small, this minimum expansion approaches $3/2$. Hence in the minimum expansion case that b or $c = 0$,

$$\text{(i) } E > 0.719\text{; (ii) } D \in (0, 1/3) \Rightarrow E > 1\text{; (iii) } D \to 0 \Rightarrow E \to 3/2. \tag{7.2}$$

In this analysis the ratio between the two kinds of difference, b and c, could be arbitrary. If the differences are produced by random noise, we might expect b and c to be approximately the same. We now consider expansion in this case that $b = c = D/2$. Since for any fixed D, $a^3 + d^3$ minimizes at $(1 - D)^3/4$ when $a = d = (1 - D)/2$, it follows that $E = (1 + 2(a^3 + d^3 + D^3 - D^2 + 3cb(1 - D)) \geq 1 + (1 - D)^3/2 + 2D^3 - 2D^2 + 3D^2(1 - D)/2 = (3 - 3D + 2D^2)/2$. This has expansion $E > 1$ for $D \in (0, 1/2)$, expansion $E = 1$ at $1/2$ and 1, and $15/16 \leq E < 1$ for $D \in (1/2, 1)$. We note that when D is small, this minimum expansion still approaches $3/2$. Hence, in the case of inputs with $b = c$, where differences in the two directions are equal, as when, for example, they are produced by noise:

$$\text{(i) } E \geq 15/16\text{; (ii) } D \in (0, 1/2) \Rightarrow E > 1\text{; (iii) } D \to 0 \Rightarrow E \to 3/2. \tag{7.3}$$

Turning now to our second equation $x + y - t \geq 1$ we get:

$t = 0$ on u, $t = 0$ on v, $U = 0$, $V = 1 : a(b^2 + 2ab)$.
$t = 0$ on u, $t = 0$ on v, $U = 1$, $V = 0 : a(c^2 + 2ac)$.

$t = 1$ on u, $t = 1$ on v, $U = 0, V = 1 : d(b^2 + 2bd)$.
$t = 1$ on u, $t = 1$ on v, $U = 1, V = 0 : d(c^2 + 2cd)$.
$t = 0$ on u, $t = 1$ on v, $U = 0, V = 1 : b(b^2)$.
$t = 0$ on u, $t = 1$ on v, $U = 1, V = 0$:
$b(1 - (a^2 + 2ab + b^2 + 2bd + d^2))$.
$t = 1$ on u, $t = 0$ on v, $U = 0, V = 1$:
$c(1 - (a^2 + 2ac + c^2 + 2cd + d^2))$.
$t = 1$ on u, $t = 0$ on v, $U = 1, V = 0 : c(c^2)$.

The sum of these eight terms is

$$X = (b + c)(1 + a^2 + d^2) - (a + d)(b^2 + c^2).$$

On substituting for the input disagreement $D = (b + c)$ and simplifying using $(a + b + c + d) = 1$, we get that the output disagreement is

$$X = D(1 + a^2 + d^2 + D^2 - D) + 2bc(1 - D).$$

Then the expansion $E = X/D = (1 + a^2 + d^2 + D^2 - D + 2bc(1 - D)/D)$. Since for any D, $a^2 + d^2$ maximizes at $(1 - D)^2$ when one of a or d equals 0 and bc maximizes at $D^2/4$ when $b = c = D/2$, it follows that E is upper-bounded by $1 + (1 - D)^2 + D^2 - D + D(1 - D)/2 = 2 - 5D/2 + 3D^2/2$. Hence for $x + y - t \geq 1$,

$$E \leq 2. \tag{7.4}$$

As far as the minimum expansion, since for any fixed D, $a^2 + d^2$ minimizes at $(1 - D)^2/2$ when $a = d = (1 - D)/2$, and bc minimizes at 0 when one of b or c is 0, it follows that $E = (1 + a^2 + d^2 + D^2 - D + 2bc(1 - D)/D) \geq 1 + (1 - D)^2/2 + D^2 - D = (3 - 4D + 3D^2)/2$. This worst-case situation, with one of b or c equal to zero, has expansion $E > 1$ for $D \in (0, 1/3)$, expansion $E = 1$ at $1/3$ and 1, and $5/6 \leq E < 1$ for $D \in (1/3, 1)$. We note that when D is small, this minimum expansion approaches 3/2. Hence, in the minimum expansion case that b or $c = 0$, exactly as before,

$$\text{(i) } E > 5/6; \text{ (ii) } D \in (0, 1/3) \Rightarrow E > 1; \text{ (iii) } D \to 0 \Rightarrow E \to 3/2. \tag{7.5}$$

8 A Construction with Arbitrarily Low Density and Inhibition

The two constructions given both converge to density $p = 0.5$. It is believed that neurons in hippocampus have an activity level corresponding to a much lower density. We now show that constructions with similar properties to the ones analyzed above also exist for lower densities.

Consider the threshold function $x + y + z - 2(t_1 + \ldots t_k) \geq 1$ over $k + 3$ variables for $k > 0$ and for k randomly chosen inputs t_j. It can be verified that this solves the problem for arbitrarily small p with $k \approx (\log_e 3)/p$. However, this construction needs the ratio of total inhibitory weights to total excitatory weights to grow linearly with $1/p$. Because there is no evidence of such an extreme ratio in hippocampus, we shall consider instead a variant that does not require it. We eliminate the need to have a significant fraction of the overall weights to be inhibitory by having a threshold function for computing t such that $t = 1$ if and only if $(t_1 + \cdots + t_k) \geq 1$, which is entirely excitatory. We use its output as a single inhibitory input to the threshold $x + y + z - 2t \geq 1$. This is the case that we shall address at length below with analysis and simulation results. (We note that, strictly speaking, this variant makes each layer of the circuit into two layers. We shall, however, refer to it as a single layer for convenience.)

As an aside, we note that one can interpolate between these two schemes, from the former where the ratio of total inhibitory synapse strength to total excitatory synapse strength grows with $1/p$, to the latter where this ratio diminishes with p. The intermediate cases would compute a threshold $x + y - 2t - 2(t_1 + \cdots + t_h) \geq 1$ where $t = 1$ if and only if $(t_{h+1} + \cdots + t_k) \geq 1$. These turn out to have properties largely independent of the value of h.

For the analysis of $x + y + z - 2t \geq 1$ where t represents $(t_1 + \ldots t_k) \geq 1$, note that the probability that the threshold is satisfied is the sum of two terms, representing the cases that $t = 0$ and $t = 1$, respectively, namely, $h(p) = (1-p)^k(1-(1-p)^3) + (1-(1-p)^k)p^3 = 3p(1-p)^{k+1} + p^3$. For a fixed point $h(p) = p$, and therefore $(1-p)^k = (1+p)/3$. For any $0 < p \leq 1/3$, there is a real number solution $k \geq 1$ since then the left-hand side is at least $2/3$ for $k = 1$, and it decreases monotonically toward 0 as k increases, while the right-hand side is in $[1/3, 2/3]$.

If $k = (\log_e 3)/p$, then the left-hand side $(1-p)^k = (1-p)^{(1/p)pk} \to e^{-log_e 3} = 1/3$. It follows that for arbitrarily small values of p, there is an integer solution k where $k = \lfloor (\log_e 3)/p \rfloor$ and our threshold function with this k has a fixed point very close to p.

For stability, we also need that $h'(p)$ at the fixed point is in the range $(-1, +1)$. It can be verified that when $k = (\log_e 3)/p$, then $h'(p) = 3(1-p)^k(1-(k+2)p) + 3p^2 \to 1 - \log_e 3 = -0.09861\ldots$ as $p \to 0$.

If we consider our other example, $x + y - 2t \geq 1$, we can again replace t by the threshold function $(t_1 + \ldots t_k) \geq 1$. Then the probability that the threshold $x + y - 2t \geq 1$ is satisfied is the sum of two terms, corresponding to the cases that $t = 0$ and $t = 1$, respectively: $h(p) = (1-p)^k(1-(1-p)^2) + (1-(1-p)^k)p^2 = 2p(1-p)^{k+1} + p^2$. For a fixed point, we need $h(p) = p$ and therefore $(1-p)^k = 1/2$. For any $0 < p \leq 1/2$, there is a real number solution $k \geq 1$ since then the left-hand side is at least $1/2$ for $k = 1$ and decreases monotonically toward 0 as k increases.

If $k = (\log_e 2)/p$, then the left-hand side $(1-p)^k = (1-p)^{(1/p)pk} \rightarrow e^{-log_e 2} = 1/2$. It follows that for arbitrarily small values of p, there is an integer solution k where $k = \lfloor (\log_e 2)/p \rfloor$ and our threshold function with this k has a fixed point very close to p.

For stability, we need that $h'(p)$ at the fixed point be in the range $(-1, +1)$. It can be verified that when $k = (\log_e 2)/p$, then $h'(p) = 2(1-p)^k(1-(k+2)p) + 2p \rightarrow 1 - \log_e 2 = 0.30685\ldots$ as $p \rightarrow 0$. Thus, the derivative is larger and hence the local rate of convergence slower for this slightly simpler family of threshold functions than those based on $x + y + z - 2t \geq 1$.

Turning to the analysis of continuity, we find that both equations inherit the desirable properties of their generating schemas $x + y + z - 2t \geq 1$ and $x + y - t \geq 1$. To see this, note that in the computation of X for the two generating schemas, we had the values a, b, c, d for the probabilities of the four conditions of the values of t in u, v, respectively. Now that t denotes the threshold function $(t_1 + \ldots t_k) \geq 1$, these four values need to be replaced by the following expressions:

$$t = 0 \text{ on } u, t = 0 \text{ on } v : a^k.$$
$$t = 0 \text{ on } u, t = 1 \text{ on } v : (a+b)^k - a^k.$$
$$t = 1 \text{ on } u, t = 0 \text{ on } v : (a+c)^k - a^k.$$
$$t = 1 \text{ on } u, t = 1 \text{ on } v : 1 + a^k - (a+b)^k - (a+c)^k.$$

It can be shown that in the limit $kb, kc \rightarrow 0$ and $a \rightarrow 1 - d$ for $x + y + z - 2t \geq 1$, the expansion is $E = 3(1-d)^k(kd + 1 - 2d) + 3d^2$, and for $x + y - t \geq 1$, the expansion is $E = 2(1-d)^k(kd + 1 - 2d) + 2d$. These give the same maximum levels of continuity as their generating schemas, namely three and two, respectively.

For orthogonality, we look to the empirical results in the next section. One complicating phenomenon is that if the stable level is $p = 0.01$, say, and one starts with a higher level of activity such as 0.025, then the amount of activity will diminish at first, and the differences b, c will also. However, as we shall see, good orthogonality is maintained by the circuit nevertheless. However, global bounds on expansion are less meaningful since low expansion is not damaging when the initial difference $D = b + c$ is large.

9 Synaptic Strengths

The constructions above show that one can get convergence in few layers, for ranges of densities larger than an order of magnitude, while maintaining continuity for arbitrarily low density probabilities, and with arbitrary ratios between excitation and inhibition. It would be interesting to determine whether one can simultaneously achieve these five requirements while also

enjoying the sixth dimension of flexibility of arbitrarily weak synapses. We have not found a general construction for this. It is possible that there are inherent limitations on achieving this sixth dimension simultaneously with the others.

However, through computer simulations reported in the next section, we have explored the space of possible parameter combinations that have smaller synaptic weights that can be still supported. In particular, we experimented with the threshold function of the form $x_1 + \cdots + x_k - (x_{k+1} + \cdots + x_{1000}) \geq C$ for various values of k and C. If $C > 1$, then the strength of individual synapses is fractional compared to the threshold, $1/C$, rather than of a magnitude at least as great as the threshold, as with our previous schemes. As an illustrative example, we show that two layers of this simple scheme for $C = 6$ have good stability in a range of densities spanning an order of magnitude, though with a higher equilibrium density than before, around 0.06, and with marginal continuity in some of the range. We do not know whether there are constructions in which this density and synaptic strengths can be arbitrarily decreased while maintaining acceptable continuity.

10 Results

We now describe the results of computer simulations of our proposed circuits as far as stability, continuity, and orthogonality.

We first note that our construction and claimed properties are such that we can obtain rigorous results by either mathematical analysis or simulations. The reason for the latter is that while the behavior of the circuits may be complex, it depends little on the actual input. For example, for stability, the expected behavior of the circuit is identical for every input having the same number of 1s. Simulations for any input will yield the same results as for any other input that has the same number of 1s. Repeating the experiment for one input with many randomly generated instances of our circuit construction will therefore reveal the behavior for all such inputs. While analysis provided the insights used to discover the circuit constructions and is applicable for whole ranges of the parameters a, b, c, and d, simulation experiments offer an equally principled way to determine the expected properties, at least for any single combination of values for a, b, c, and d.

In reporting the results, we shall describe how they relate to the previously described analysis. The results we report first are for the threshold $x + y + z - 2t \geq 1$ when t has the value 1 or 0 according to whether $(t_1 + \cdots + t_{109}) \geq 1$ holds. The value of $k = 109$ was used so as to approximate the equilibrium density $p = 0.01$. (Simulations for $x + y - t \geq 1$ with $k = 69$ yielded similar results but with stability in a narrower range, though with better continuity.)

For each combination of parameters, the mean values of the properties to be estimated were computed by taking the means over 100 runs, where each run consisted of constructing new random connections for the whole

network and applying it to an input with the appropriate parameters a, b, c, d. In the simulations, 1 million neurons were used at each layer of the network. We estimate continuity by simulating pairs of inputs with equal numbers of 0s and 1s, so that the differences between them are balanced between the 0s and 1s. This was the $c = b$ case at which maximum expansion is achieved for one layer of the underlying threshold. We estimate orthogonality by simulating pairs of inputs with $c = 0$, so that $bc = 0$, when the minimum expansion is achieved for one layer for the underlying threshold.

The results can be summarized as follows. In the range of input densities $[q, s] = [0.002, 0.025]$, the construction based on $x + y + z - 2t \geq 1$, with three layers, behaves well in all of the following three senses:

1. It is 0.01-stable in that for any input, the construction will produce an output with mean within 1% of the equilibrium level of $p = 0.01$, namely, in the range $[0.0099, 0.0101]$.
2. For any two inputs differing in a small fraction $x = b + c$ of bits and with $b = c$, the outputs after three layers differ by at most a fraction $18x$ in expectation. This maximum expansion was achieved in this range at the minimum density 0.002. (The expansion was at most $10x$ when the density is the mean of $p = 0.01$.) This level of expansion ensures, for example, that if items represented by fraction 0.01 of 1s are regarded as the same by cortex whenever their difference is less than fraction 0.001 (i.e., 10% of the density), then an error rate of 1/18 of this latter quantity, fraction 0.000505.., in the firing probability of the input neurons can be tolerated.
3. For any two inputs that differ in a fraction y of the bits with $c = 0$ the outputs differed by at least a fraction $0.93y$ in expectation throughout the range.

Table 1 shows the densities reached after each of the first four layers of our circuit for the threshold $x + y + z - 2t \geq 1$ with $k = 109$. Our analysis predicted that the equilibrium value is the solution to the equation $(1 - p)^{109} = (1 + p)/3$, which is $p = 0.0099953\ldots$. (For the second threshold $x + y - t \geq 1$, the choice $k = 69$ gives the predicted equilibrium value to be the solution to $(1 - p)^k = 1/2$, which is $p = 0.0099385\ldots$)

Figure 1 shows the expansion after layer 3 of the circuit for input v with density $b+d$ at the extreme values of 0.002 and 0.025 and at the equilibrium value of 0.01. In each case, we give the expansion achieved at $b = c$ (the continuity estimate) and the expansion achieved at $c = 0$ (the orthogonality estimate). These six values are given for a range of values of the input Hamming distance $b + c$. Note that for any v density $b + d$, the densities of u will be different for the orthogonality and continuity estimates, which therefore are not pairwise comparable.

We observe that the continuity values are consistent with the analysis. In the limit that b, c are small and hence the input density approaches d, the analysis predicts expansion $3(1 - d)^k(kd + 1 - 2d) + 3d$. Since this is

Table 1: Output Densities Achieved at Various Depths.

Input	Level 1	Level 2	Level 3	Level 4
0.0400	0.00135	0.00348	0.00713	0.00974
0.0300	0.00315	0.00667	0.00958	0.00997
0.0250	0.00464	0.00834	0.00996	0.00994
0.0200	0.00650	0.00950	0.00997	0.00993
0.0150	0.00854	0.00996	0.00995	0.00993
0.0100	0.00992	0.00995	0.00995	0.00993
0.0075	0.00983	0.00996	0.00992	0.00993
0.0050	0.00865	0.01000	0.00992	0.00995
0.0033	0.00690	0.00967	0.00996	0.00994
0.0020	0.00482	0.00849	0.00996	0.00993
0.0015	0.00383	0.00754	0.00984	0.00994
0.0010	0.00271	0.00603	0.00929	0.00999

Notes: For a range of input densities between 0.001 and 0.04 the mean densities of the outputs of the circuit after layers 1, 2, 3, and 4, respectively, are shown. The entries are means over 100 simulations. Note that within the range [0.002, 0.025] of input densities, the mean density after the third layer is within 1% of the equilibrium value 0.01. Within that range, the standard deviations measured were all in the range [0.00008, 0.00012]. With a less stringent tolerance requirement, 5%, stability is reached after two layers within the narrower range [0.0033, 0.02], while after four levels, 1% stability is achieved in the broader range [0.001, 0.03].

upper-bounded by 3, an upper bound of 27 follows for the continuity of the whole circuit. However, the contribution of each layer is less than 3 since d is nonzero and depends on the actual density achieved at that layer. For example, substituting $k = 109$ gives an expansion of $E = 2.10\ldots$ for density $d = 0.01$. At the higher density of $d = 0.025$, we have a smaller expansion $E = 0.77\ldots$, and at the lower density of $d = 0.002$, we have a higher expansion $E = 2.93\ldots$ For small enough b, c, continuity and orthogonality are therefore the same. Examination of the expansion shown at intermediate layers in the simulation reveals that the measures for continuity and orthogonality are sometimes significantly different after one layer but converge by the third.

We also simulated a number of variants that produced essentially the same results as shown in Table 1 and Figure 1. First, we found that the threshold $x + y + z - 2t - 2(t_1 + \cdots + t_h) \geq 1$ where $t = 1$ if and only if $(t_{h+1} + \cdots + t_k) \geq 1$ for $k = 109$ gives essentially the same results for any value of $h (0 \leq h \leq k)$. This permits the inhibitory-excitatory ratio to be anywhere in the range $1/p$ and p depending on the value of h.

Second, we found that the circuit does not need full randomization. For example, if the million-neuron wide circuit is split into 100 independent circuits, each of one-hundredth of the width, one gets essentially the same results as long as the 1s in the input are divided about equally among the

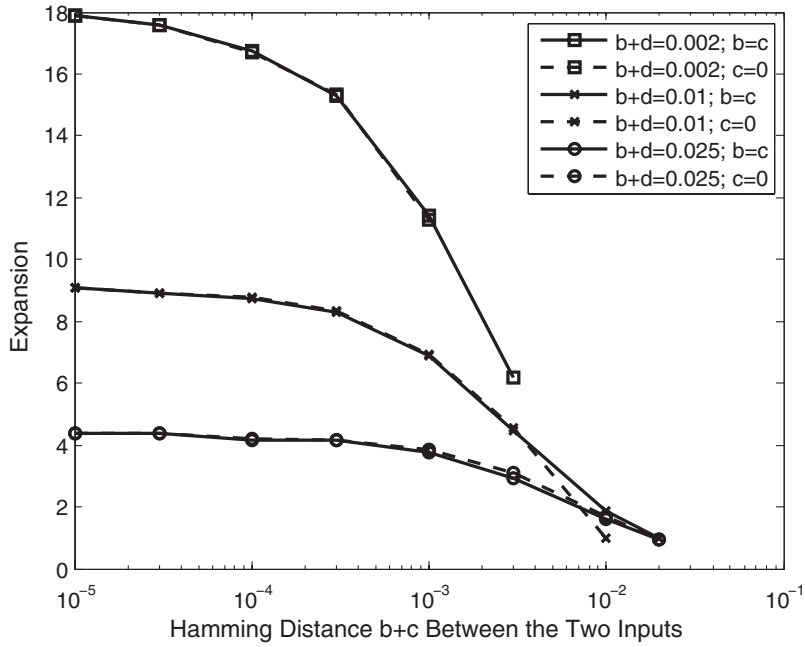

Figure 1: The expansion achieved after layer 3 of the circuit for inputs u, v for six different density combinations, each for up to eight different values of the input Hamming distance b+c in the range [0.00001, 0.02]. Three of the density combinations have $b = c$ and are for $Dense(u) = Dense(v) = b + d$ equal to 0.002, 0.01, and 0.025. The other three have density combinations $c = 0$ and are for Dense$(v) = b + d$ equal to 0.002, 0.01, and 0.025. Note that in either case, it is not meaningful to have the input Hamming distance b+c larger than twice the input density b+d; hence, the corresponding cases are omitted. Also note that the outlier combination $b + c = 0.01$, $c = 0$ with Dense$(v) = b + d = 0.01$ trivializes in that then $d = 0$ and hence u is the zero vector. The cases $b = c$ and $c = 0$ were not found otherwise to be significantly different after three layers, though they sometimes were after one or two. Note that for any one value of Dense$(v) = b + d$, the two estimates are not strictly comparable with each other since the Dense(u) differ.

100 subcircuits. This shows that substantial locality of connectivity within the SMA can be tolerated and global randomization is not required. It also shows that circuits of width 10^4 perform similar to those with 10^6 except for having higher variances.

Finally we simulated the scheme with fractional synaptic weights, based on the threshold $x_1 + \cdots + x_k - (x_{k+1} + \cdots + x_{1000}) \geq 6$. We averaged over 100 simulations and with each layer having $200,000$ neurons. The results are shown in Figure 2. Stability within 10% of the equilibrium

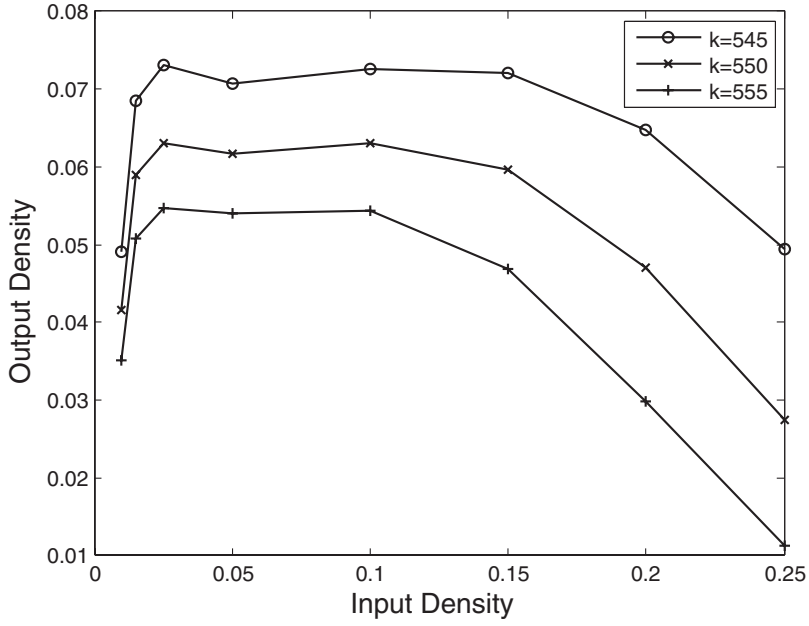

Figure 2: The densities after two layers for the threshold $x_1 + \cdots + x_k - (x_{k+1} + \cdots + x_{1000}) \geq 6$ for the three values $k = 545$, 550, and 555. In each of the three cases, the output density lies within 10% of the equilibrium for the range of input densities [0.015, 0.15] spanning an order of magnitude. Within that range in all cases, the expansion was found to be in the range 99 to 417, the maximum being reached for $d = 0.015$ and $k = 545$ and the minimum for $d = 0.15$ and $k = 555$. The results were very similar for the four cases with $b + c = 0.0001$ or 0.00001 and $b = c$ or $c = 0$.

density is achieved for an order of magnitude range [0.015, 0.15] of input densities. Among the three values of k, the minimum expansion was obtained for $k = 555$, and for this, the output density difference x when the input density difference was 0.00001 was as follows when expressed as (*inputdensity*, x) pairs: (0.015, 0.0041), (0.025, 0.0039), (0.05, 0.0029), (0.1, 0.0022), (0.15, 0.0017). From this, we see that at the low-density end, the output difference will be 27% of the density, which is too high for acceptable continuity. Hence, to tolerate this level of continuity at the less dense end of this range, the errors that can be tolerated will be smaller than 0.00001. While our main construction can simultaneously achieve arbitrarily low equilibrium densities and acceptable continuity, it remains an open problem whether the same holds for any construction with small fractional synaptic weights.

11 The Hippocampus

11.1 Architecture of the Hippocampus. The overall flow of information within the hippocampal system has a strong unidirectional aspect, starting from the entorhinal cortex, successively though the dentate gyrus, CA3, CA2, CA1, the subiculum, and back to the entorhinal cortex. There are also additional connections between these regions that are in the same direction and bypass intermediate regions. In contrast to the reciprocal connectivity widely found in the rest of the mammalian brain, reverse connections among these hippocampal regions appear to be fewer (Anderson, Morris, Amaral, Bliss, & O'Keefe, 2007). In this letter, we have shown that SMA functionality can be achieved in feedforward networks having as few as two or three levels.

Estimated neuron numbers for humans (West & Gunderson, 1990; Simic, Kostovic, Winblad, & Bogdanovic, 1997; Walker et al., 2002) are 13 to 18 million in dentate gyrus and 2 to 3 million in CA3 and 12 to 21 million in CA1. The described simulation results are for 1 million neurons at each level. Simulations yield essentially the same results with numbers that are higher, or as low 10,000 except in the latter case the variance increases.

Within some areas of hippocampus, there are additional connections that suggest a more recurrent aspect internal to these areas, particularly in CA3. These may be necessary for functions that we do not analyze here. In particular, it is widely held that hippocampus has to consolidate the representation in cortex over a period of weeks or longer. Simultaneous recordings from cortex and hippocampus during sleep have yielded direct support for this (Wierzynski, Lubenov, Gu, & Siapas, 2009). Hence, recurrent connections in cortex may be being used to store the information necessary to effect the consolidation process over such longer periods.

As far as information flow to and from the hippocampal system, there is believed to be reciprocal flow to many neocortical association areas, mainly via the entorhinal cortex. This is consistent with our hypothesis that hippocampus is a memory allocator for cortex in general.

11.2 Randomness of Connections in Hippocampus. Our thesis that the hippocampus has effectively random connectivity between the levels is consistent with evidence that the firing patterns of anatomically close pyramidal neurons in hippocampus are uncorrelated (Redish et al., 2001). It is also consistent with evidence that the place fields of anatomically close place cells are similarly uncorrelated (Thompson & Best, 1990). Some evidence for the contrary has been also reported (Hampson, 1999).

11.3 Activity Level in Hippocampus. What activity level does the hippocampus exhibit that would correspond to our density parameter p? From the results of recordings of medial temporal lobe of humans presented with visual scenes, various estimates have been made, including ones of 0.0054

and 0.012 (Waydo, Kraskov, Quiroga, Fried, & Koch, 2006). We have chosen the value of 0.01 as illustrative, but our methods work equally well for any value. We also note that the activity level in cortex may be totally different from that in hippocampus. It is easily verified that an SMA working at one density can interface with cortex working at another via an appropriate one-layer bipartite network that translates the one density to the other.

11.4 Permanence of Function of Hippocampus. There is evidence that place cells have substantial permanence. Thompson and Best (1990) have shown that for periods of up to 153 days, the place fields of place cells in rat hippocampus are highly stable. Our SMA theory does require such permanence for the period during which individual chunks are consolidated (but see the comments below about neurogenesis.)

11.5 Unmodifiable Synapses. We note that the basic SMA functionality that we ascribe to the hippocampus involves no learning and, in fact, requires that the relevant synaptic weights not change during the cortical consolidation process. Evidence has been reported that there exist synapses in the hippocampus that are not modifiable (Peterson, Malenka, Nicoll, & Hopfield, 1998; Debanne, Gahwiler, & Thompson, 1999). This is consistent with our theory but not essential to it.

11.6 The Role of Modifiable Synapses. Modifiable synapses abound in hippocampus, and the question arises as to the function of these. As mentioned above, it is widely believed that memories are consolidated by hippocampus over a longer period. In our framework, the hippocampus stimulates the neurons where the chunk $A\&B$ is to be stored when the constituents A, B are firing. Hence, the hippocampus would need to retain some information, such as the firing patterns of A, B, that would be necessary to complete the consolidation. It is reasonable to believe that this information is stored in hippocampus in the form of modifiable synapses.

11.7 Bigger Chunks. Our conjecture is that hippocampus can perform memory allocation for conjunctions of r items—not just for some fixed size such as $r = 2$, but in a range of sizes $2 \leq r \leq R$, in such a way that the number of neurons allocated is about the same whether $r = 2$ or $r = 5$, say. A major drawback of the alternative known approaches to stability that we reviewed earlier is that they appear to be specific to memorization of conjunctions of some fixed size, such as two, rather than a range of sizes.

11.8 Neurogenesis. It is believed that in the dentate gyrus, new neurons appear at a steady rate even in adulthood, though at a rate decreasing with age. Cameron and McKay (2001) showed that in young adult rats, about 3600 new granule cells were produced each day in dentate gyrus, among a total of $1.5 - 2 \times 10^6$ such cells. It is believed that the cell death rate is

approximately equal, which implies a turnover rate of 5% to 7% per month. The deficits in learning that the prevention of neurogenesis in dentate gyrus produces have been also explored but with few definitive conclusions to date (Anderson et al., 2007).

What purpose does neurogenesis achieve in an SMA? The effect of neurogenesis will be to change the addresses computed by the SMA over time for a fixed chunk. One interpretation of this is that the SMA adds a time stamp to the allocation, so that the same image a year later will cause a new episodic memory to be made. Within the time needed for memory consolidation, the effect of neurogenesis has to be small enough that by continuity, the output changes little. With this interpretation, the fact that neurogenesis happens in the earliest layer, the dentate gyrus, suggests that the largest rate of functional change is being achieved per new neurons created. If neurogenesis is interpreted as the replacement of some input neurons by some new neurons, with new random connections to the first layer of the feedforward circuit, then the effect of neurogenesis can be analyzed in the same terms as continuity and orthogonality.

If the effect of neurogenesis is to change the functions computed by the SMA at a steady rate, the question arises whether neurogenesis is the best way of effecting this result. Randomly changing the connections within the SMA at a steady rate would have a similar effect.

11.9 Synaptic Strengths. The distributions of synaptic strengths in hippocampus are not well understood, nor is the exact relationship between the measured strength of single synapses and the effect of that synapse in vivo. However, synapses in hippocampus that are strong enough that action potentials in the presynaptic neuron can cause action potentials in the postsynaptic neuron have been reported (Miles & Wong, 1983, 1986; Csicsvari, Hirase, Czurko, & Buzsaki, 1990, 1998). Our main constructions assume this. We have also given an example of a weak synapse construction, where six presynaptic neurons are needed to cause action potentials. However, it remains an open question as to whether SMAs with arbitrarily low density and good continuity can be realized with arbitrarily weak synapses.

11.10 Nature and Amount of Inhibition. The question of whether inhibition is subtractive or divisive has been widely discussed in the literature (Abbott & Chance, 2005). We note that threshold functions we consider are of the subtractive form $X - T \geq 1$ but can be equivalently replaced by the divisive or shunting inhibition version $X/(X + T) > 1/2$, which is equivalent to $X - T > 0$, which in turn is equivalent to the original $X - T \geq 1$ if X and T are integers.

We have pointed out that our constructions have variants where the ratio of total inhibitory synapse strength to total excitatory synapse strength grows with $1/p$, diminishes with $1/p$, or is anywhere in between these extremes. This ratio differs for different cell types. For pyramidal cells in

CA1, these two kinds of synapses have been estimated to number 1,700 and 30,000, respectively (Megias, Emri, Freund, & Gulyas, 2001).

12 Conclusion

We have considered hierarchical representations in cortex where an item or chunk at one level is a conjunction of items or chunks at other levels. We have given a construction for feedforward circuits with three layers that are able to allocate neurons to new chunks so that the number of neurons needed for any chunk will be close to an expected value. The construction is highly flexible, being (1) stable within a 1% range after just three layers throughout an order of magnitude range of input densities, (2) valid for any number of neurons above 10,000, (3) tolerant to widely different ratios between inhibitory and excitatory connections, (4) resistant to noise, and (5) adaptable to any density of activity—the illustrative value of $p = 0.01$ having been chosen as a typical value consistent with experiments. Our main construction requires that synapses be strong enough that single neurons firing at one level have a significant influence on the neurons at the next level. While there is experimental evidence for such influences, we also report on stable circuit schemes that require only weaker synapses. These schemes, however, are viable in more limited parameter ranges, and it is not known whether arbitrarily low activity and synaptic influence levels are consistent with noise tolerance.

The hippocampus clearly has additional functions that we have not addressed. For example, besides identifying neurons in cortex, it also needs to store information to be used when consolidating memories at those neurons over a period. However, we believe that the memory allocation function that we attribute here to hippocampus corresponds well with the main function that the mammalian brain with an impaired hippocampus appears to lack.

Independent of this, our construction solves a nagging theoretical problem in the neuroidal model. Although the neuroidal model has been shown to be able to support a useful variety of functions, more than any other model, we believe it was missing a mechanism that guarantees the stability of hierarchical memory allocation. The task of memory allocation, for which we could not find a plausible cortical mechanism, is, however, very reminiscent of the task that individuals with hippocampal damage cannot perform. Hence, the hypothesis that this task is in fact carried out by the hippocampus is a natural one to make.

An interesting next step would be to incorporate the proposed stable memory allocator in a significant simulation of cortical functions. In the simulation of Feldman and Valiant (2009), memory allocation consists of assigning primitive items to a fixed number of randomly chosen neurons and one level of further allocation by applying JOIN to pairs of these primitive items. (Applying a second level of JOIN had the effect of considerably

degrading the other operations.) Repeating these simulations, but with arbitrary depth of memory allocation using the SMA construction of this letter, would be valuable. One would imagine that these earlier results could be reproduced, since the new items allocated using the SMA would have similar pseudorandom properties of having small intersections and similar sizes, as do the first-level allocations by JOIN that were tested. One can control the quality of the SMA by varying its depth. It would be interesting to demonstrate that low- or moderate-quality SMAs are sufficient to support the cortical functions being tested.

Cortex hosts an ocean of computational processes, the success of each of which is, we believe, guaranteed only by statistical factors. We are suggesting here that for the cortex to succeed in this imprecise regime, it needs the help of a precision instrument and that the hippocampus is that instrument. The mechanism we propose for the hippocampus requires only the same imprecise statistical regime as cortex, but since its function is specialized to memory allocation, precision of behavior is, as we have demonstrated here, possible nevertheless.

Acknowledgments

I am grateful to Kenny Blum, Joachim Diederich, Gabriel Kreiman, John O'Keefe, Richard Miles, Dan Schacter, and Steven Siegelbaum for variously answering questions and bringing references to my attention. I am also grateful to Peter Dayan, Howard Eichenbaum, Scott Linderman, Haim Sompolinsky, and two anonymous referees for some most helpful comments on this manuscript. This research was partly funded by NSF grants CCF-04-27129 and CCF-09-64401.

References

Abbott, L. F., & Chance, F. S. (2005). Drivers and modulators from push-pull and balanced synaptic input. *Progress in Brain Research*, *149*, 147–155.

Amari, S. (1974). A method of statistical neurodynamics. *Kybernetik*, *14*, 201–215.

Amit, D. J., & Brunel, N. (1997). Dynamics of a recurrent network of spiking neurons before and following learning. *Computation in Neural Systems*, *8*, 343–404.

Anderson, P., Morris, R., Amaral, D., Bliss, T., & O'Keefe, J. (2007). *The hippocampus book*. New York: Oxford University Press.

Beal, J., & Knight, T. F. (2008). Analyzing composability in a sparse encoding model of memorization and association. In *Proc. 7th IEEE International Conference on Development and Learning (ICDL 2008)* (pp. 180–185). Piscataway, NJ: IEEE.

Cameron, H. A., & McKay, R.D.G. (2001). Adult neurogenesis produces a large pool of new granule cells in the dentate gyrus. *J. Comp. Neurology*, *435*, 406–417.

Cohen, N. J. (1981). *Neuropsychological evidence for a distinction betwen procedural and declarative knowledge in human memory and amnesis*. Unpublished doctoral dissertation, University of California at San Diego.

Cohen, N. J., & Eichenbaum, H. (1993). *Memory, amnesia and the hippocampal system*. Cambridge, MA: MIT Press.

Csicsvari, J., Hirase, H., Czurko, A., & Buzsaki, G. (1990). Synaptic excitation of inhibitory cells by single CA3 hippocampal pyramidal cells of the guinea-pig in vitro. *J. Physiol.*, *428*, 61–77.

Csicsvari, J., Hirase, H., Czurko, A., & Buzsaki, G. (1998). Reliability and state dependence of pyramidal cell interneuron synapses in the hippocampus: An ensemble approach in the behaving rat. *Neuron*, *21*, 179–189.

Debanne, D., Gahwiler, B. H., & Thompson, S. M. (1999). Heterogeneity of synaptic plasticity at unitary CA3-CA1 and CA3-CA3 connections in rat hippocampal slice cultures. *J. Neurosci.*, *19*, 10664–10671.

Diederich, J., Gunay, C., & Hogan, J. M. (2010). *Recruitment learning*. Berlin: Springer.

Feldman, J. A. (1982). Dynamic connections in neural networks. *Biol. Cybern.*, *46*, 27–39.

Feldman, V., & Valiant, L. G. (2009). Experience induced neural circuits that achieve high capacity. *Neural Computation*, *21*, 2715–2754.

Gerbessiotis, A. V. (2003). Random graphs in a neural computation model. *International Journal of Computer Mathematics*, *80*, 689–707.

Graham, B., & Willshaw, D. (1997). Capacity and information efficiency of the associative net. *Network: Comput. Neural Syst.*, *8*, 35–54.

Gunay, C., & Maida, A. S. (2006). A stochastic population approach to the problem of stable recruitment hierarchies in spiking neural networks. *Biol. Cybern.*, *94*(1), 33–45.

Hampson, R. E. (1999). Distribution of spatial and nonspatial information in dorsal hippocampus. *Nature*, *402*, 610–614.

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

Kali, S., & Dayan, P. (2004). Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nature Neuroscience*, *7*, 286–294.

Latham, P. E., & Nirenberg, S. (2004). Computing and stability in cortical networks. *Neural Computation*, *16*, 1385–1412.

Marr, D. (1971). Simple memory: A theory of archicortex. *Phil. Trans. Roy. Soc. B*, *262*, 23–81.

Martin, V. C., Schacter, D. L., Collins, M. C., & Rose, D. R. (2011). A role for the hippocampus in encoding simulations of future events. *Proc. Natl. Acad. Sci.*, *108*(33), 13858–13863.

Megias, M., Emri, Z., Freund, T. F., & Gulyas, A. I. (2001). Total number and distribution of inhibitory and excitatory synapses on hippocampal CA1 pyramidal cells. *Neuroscience*, *102*, 527–540.

Miles, R., & Wong, R. K. S. (1983). Single neurones can initiate synchronized population discharge in the hippocampus. *Nature*, *306*, 371–373.

Miles, R., & Wong, R. K. S. (1986). Excitatory synaptic interactions between CA3 neurones in the guinea-pig hippocampus. *J. Physiol*, *373*, 397–418.

Minai, A. A., & Levy, W. B. (1994). Setting the activity level in sparse random networks. *Neural Computation*, *6*, 85–99.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: Clarendon Press.

Peterson, C. C. H., Malenka, R. C., Nicoll, R. A., & Hopfield, J. J. (1998). All-or-none potentiation at CA3-CA1 synapses. *Proc. Natl. Acad. Sci.*, *95*, 4732–4737.

Redish, A. D., Battaglia, F. P., Chawla, M. K., Ekstrom, A. D., Gerard, J. L., Lipa, P., et al. (2001). Independence of firing correlates of anatomically proximate hippocampal pyramidal cells. *J. Neuroscience*, *21*, RC134.

Rolls, E. T. (1996). A theory of hippocampal function in memory. *Hippocampus*, *6*, 601–620.

Schacter, D. L., & Buckner, R. L. (1998). Priming and the brain. *Neuron*, *20*, 185–195.

Schacter, D., & Tulving, E. (1994). *Memory systems*. Cambridge, MA: MIT Press.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatry*, *20*, 11–21.

Simic, G., Kostovic, I., Winblad, B., & Bogdanovic, N. (1997). Volume and number of neurons of the human hippocampal formation in normal aging and Alzheimer's disease. *J. of Comparative Neurology*, *379*, 482–494.

Smith, A. C., Wu, X. B., & Levy, W. B. (2006). Controlling activity fluctuations in large, sparse connected random networks. *Network: Comput. Neural Syst.*, *11*, 63–81.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys and humans. *Psychol. Rev.*, *99*, 195–231.

Tegne, J., Compte, A., & Wang, X.-J. (2002). The dynamical stability of reverberatory neural circuits. *Biol. Cybern.*, *87*, 471–481.

Teyler, T. J., & DiScenna, P. (1986). The hippocampal memory indexing theory. *Behav. Neurosci.*, *100*, 147–152.

Thompson, L. T., & Best, P. J. (1990). Long-term stability of the place-field activity of single units recorded from the dorsal hippocampus of freely behaving rats. *Brain Research*, *509*, 299–308.

Tulving, E. (1983). *Elements of episodic memory*. Oxford: Clarendon Press.

Valiant, L. G. (1994). *Circuits of the mind*. New York: Oxford University Press.

Valiant, L. G. (2005). Memorization and association on a realistic neural model. *Neural Computation*, *17*, 527–555.

Walker, M. A., Highley, J. R., Esiri, M. M., McDonald, B., Roberts, H. C., Evans, S. P., et al. (2002). Estimated neuronal populations and volumes of the hippocampus and its subfields in schizophrenia. *Am. J. Psychiatry*, *159*, 821–828.

Waydo, S., Kraskov, A., Quiroga, R. Q., Fried, I., & Koch, C. (2006). Sparse representation in the human medial temporal lobe. *Journal of Neuroscience*, *26*, 10232–10234.

West, M. J., & Gunderson, H.J.G. (1990). Unbiased stereological estimation of the number of neurons in the human hippocampus. *Journal of Comparative Neurology*, *296*, 1–22.

Wickelgren, W. A. (1979). Chunking and consolidation. *Psychological Review*, *86*, 44–60.

Wierzynski, C. M., Lubenov, E. V., Gu, M., & Siapas, A. G. (2009). State-dependent spike timing relationships between hippocampal and prefrontal circuits during sleep. *Neuron*, *61*(4), 587–596.

Received February 24, 2012; accepted June 1, 2012.