Statistical implications of clipped Hebbian learning of cell assemblies. [1]

Andreas Knoblauch [a]

[a] *Department of Neural Information Processing*
University of Ulm, Oberer Eselsberg, D-89069 Ulm, Germany
email: knoblauch@neuro.informatik.uni-ulm.de

Abstract

Although cell assemblies have been postulated by Donald Hebb almost half a century ago, so far they have not yet been proven (or disproven) to occur in the real brain. This is mainly because of immense difficulties in recording simultaneously from a large number of single neurons with high spatial and temporal resolution. In this study I suggest an alternative approach to test the structure of a local cortical network. After repeated stimulation of a large neuron number, the test just requires the evaluation of statistical properties of the postsynaptic potentials recorded from a single cell. Using a simple binary network model and applying clipped Hebbian learning, it is shown that the variance in the postsynaptic potentials grows with the square of the stimulation strength if the synapses have been generated by Hebbian learning of many overlapping cell assemblies, but only linearly for independent random synapses. This result bears implications both for analysis of associative memory and the verification of the assembly hypothesis in neurophysiological experiments.

Key words: cell assemblies, Hebbian learning, associative memory, Willshaw model, postsynaptic potentials

1 Introduction

Although cell assemblies have been postulated by Donald Hebb almost half a century ago [4], so far they have not yet been proven (or disproven) to occur in the real brain. This is mainly because of immense difficulties in recording simultaneously from a large number of single neurons with high spatial and temporal resolution. In this study I introduce an alternative approach to test

[1] The author has been supported by the MirrorBot project of the European Union.

the local network by first stimulating a large number of neurons but then evaluating the statistical properties of the postsynaptic potentials recorded from only a single cell that is directly connected to the stimulated population. Using a simple binary network model it is shown that the variance in the postsynaptic potentials grows with the square of the stimulation strength if the synapses have been generated by Hebbian learning of many overlapping cell assemblies, but only linearly for independently generated random synapses.

2 Binary associative networks

Binary associative networks have been introduced as a very abstract model for cell assemblies in the local cortical connections [13,14,8–10,1]. Figure 1 illustrates how two patterns (or synonymously, assemblies or attractors) of length $n = 7$ each containing $k = 4$ one-entries are stored autoassociatively in the synapses of a local neuron population by Hebbian learning. If a binary pattern u^μ is active then the strengths of all k^2 synapses connecting two active neurons are increased (from 0 to 1). When learning another pattern then increasing the strength of an already activated synapse has no further effect (*clipped* Hebbian learning). Thus the synapses are binary (either 0 or 1), and the assemblies can be interpreted as fully connected subsets of the neurons (i.e., as k-cliques of a graph with n nodes). After learning of M patterns the strength of the synaptic connection from neuron i to neuron j can be written as

$$A_{ij} = \min\left(1, \sum_{\mu=1}^{M} u_i^\mu \cdot u_j^\mu\right) \in \{0, 1\}\,. \tag{1}$$

For pattern *retrieval* we address the network using an address pattern $\tilde{u}$ which may (for example) be a noisy version of one of the original address patterns, u^μ. By vector-matrix-multiplication we obtain the neural potentials $x = A \cdot \tilde{u}$ which can be transformed to the retrieval result $\hat{u}$ by applying a threshold Θ, i.e. $\hat{u}_i = 1$ if $x_i \geq \Theta$ and 0 otherwise for $i = 1, ..., n$. The choice of the threshold Θ is important for good retrieval results. One possibility which is referred to as the *Willshaw-retrieval-strategy* is simply setting the threshold equal to the number of one-entries in the address pattern $\tilde{u}$, $\Theta = \sum_{i=1}^{n} \tilde{u}_i$. If all the one-entries in the address pattern $\tilde{u}$ occur also in one of the original patterns, u^μ, then the one-entries in the retrieval result $\hat{u}$ will always be a superset of the ones in the original pattern. Indeed, this strategy is the only possible choice if one assumes that the address pattern contains no 'false alarms', and it plays also an important role for pattern separation in spiking associative memories with time-continuous retrievals (cf. [7,6]).

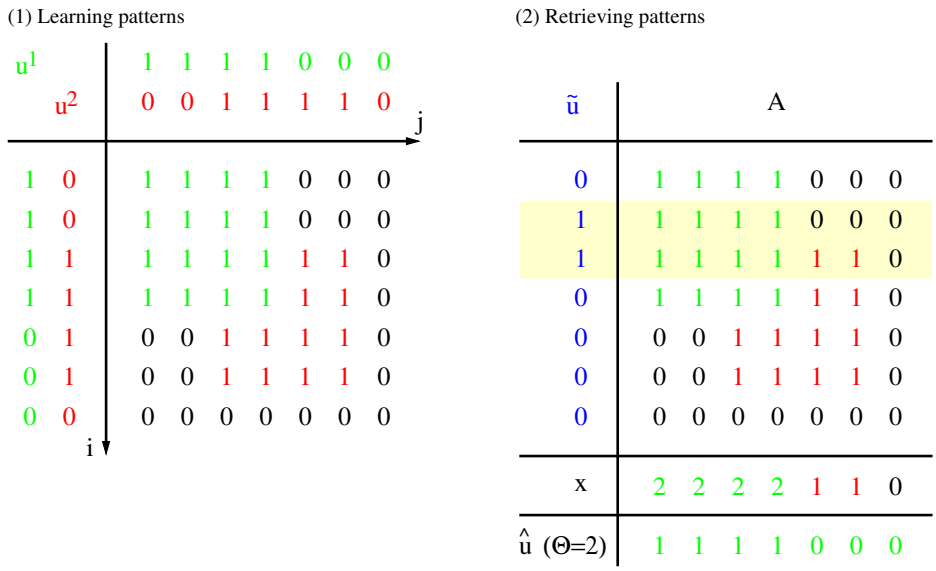

Fig. 1. Learning and retrieving patterns in a binary auto-associative network.

Generally, the probability of a retrieval error will increase with the fraction p_1 of active synapses, which is also referred to as the *matrix load* p_1 which increases with the number M of stored patterns. For random patterns we obtain

$$p_1 \approx 1 - (1 - k^2/n^2)^M \approx 1 - e^{-Mk^2/n^2}, \tag{2}$$

It is the main matter of the theory of neural associative memory to determine how much information and how many patterns can be stored safely in a network of n neurons [8,11,5,6]. The general strategy is to divide the neuron population into "correct" and "false" neurons, and then to determine the two distributions of membrane potentials for the two classes. Finally, applying an optimally chosen threshold Θ extracts ideally exactly the "correct" neurons.

A popular approximation is to assume that the activated synapses would have been generated independently of each other which neglects the fact that storing a pattern will activate a block of k^2 synapses at a time. However, applying this *binomial approximation*, analysis becomes relatively easy. For example, let us assume that the address pattern $\tilde{u}$ contains $\lambda \cdot k$ one-entries of a previously stored pattern u^μ, and $\kappa \cdot k$ randomly chosen noisy one-entries ($0 < \lambda \leq 1$, $\kappa > 0$), such that the total address pattern activity is $z := (\lambda + \kappa) \cdot k$. Then using the binomial probability,

$$p_B(x; N, p) := \binom{N}{x} \cdot p^x \cdot (1-p)^{N-x}, \tag{3}$$

the membrane potential X_c of a "correct" neuron, and the membrane potential

X_f of a "false" neuron are distributed according to the following probabilities,

$$\mathrm{pr}[X_c = x] = p_B(x - \lambda \cdot k; \kappa \cdot k, p_1), \tag{4}$$
$$\mathrm{pr}[X_f = x] = p_B(x; z, p_1). \tag{5}$$

For $\kappa = 0$, for example, we can simply apply the Willshaw threshold strategy ($\Theta = z = \lambda \cdot k$) and write for the error probability that a given "false" neuron gets activated, $p_{01} = \mathrm{pr}[X_f = z] \approx p_B(z; z, p_1) = {p_1}^z$.

Unfortunately, it turns out that the binomial approximation can be quite bad, both qualitatively and quantitatively [6]. Figure 2 illustrates two phenomena that can occur for the potential distributions: (1) Oscillatory modulations and (2) massive underestimation of the variance, in particular for large address pattern activity z.

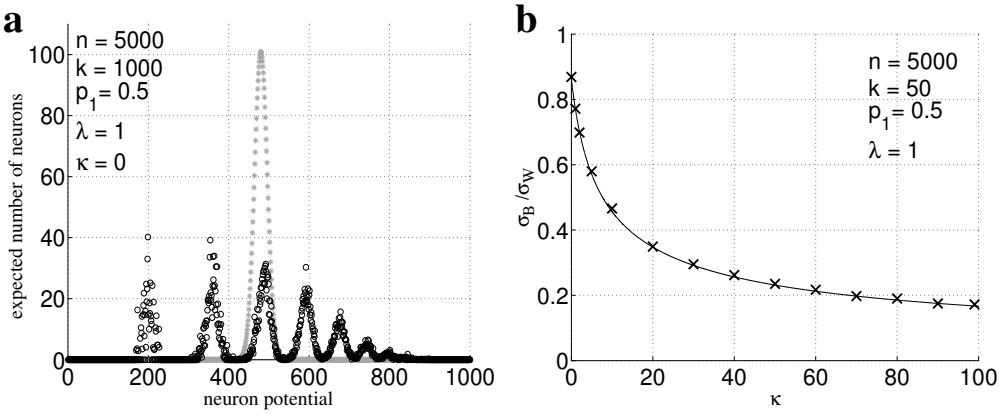

Fig. 2. The binomial approximation of the neuron potential distribution can be bad. **a**: Simulations reveal oscillatory modulations in the neuron potentials (black) not captured by the binomial approximation (gray). **b**: The s.d. σ_B of the binomial approximation can underestimate massively the true s.d. σ_W of the Willshaw distribution. *Crosses* refer to simulation experiments, *solid line* uses eqs. 14 and 15 (cf. [6]).

Both phenomena can be explained by the exact formula $\mathrm{pr}[X_f = x] = p_W(x; n, k, M, z)$ found by Buckingham and Willshaw ([3,2]; see also [12] and the appendix in [8]),

$$p_W(x; n, k, M, z) := \sum_{i=0}^{M} p_B(i; M, \frac{k}{n}) \cdot p_B(x; z, 1 - (1 - \frac{k}{n})^i)., \tag{6}$$

which, however, cannot be evaluated and analyzed as easily as the binomial approximation. In the following p_W is referred to as the *Willshaw distribution*. First note that p_W is a superposition of M binomial probabilities p_B. This explains the oscillatory modulations which occur if the standard deviation (s.d.) of the component binomials is small compared to the distances of the

means (see [6] for details). In the following section I will determine a good approximation for the true variance of p_W which explains the underestimation by the binomial approximation.

3 The variance of the Willshaw distribution

Now we try to approximate the variance $\mathrm{Var}(X_R) = E({X_R}^2) - E(X_R)^2$ from the formula of Buckingham and Willshaw (eq. 6). First note that the binomial approximation already gives us the correct expectation (cf. eq. 2)

$$E(X_R) = z \cdot p_1 = z \cdot (1 - (1 - \frac{k^2}{n^2})^M) \tag{7}$$

In order to infer the variance of the Willshaw distribution from eq. 6 the following equations turn out to be useful,

$$\sum_{i=0}^{M} p_B(i; M, p) \cdot (1-p)^i \;=\; (1-p^2)^M \qquad (\approx e^{-Mp^2}), \tag{8}$$

$$\sum_{i=0}^{M} p_B(i; M, p) \cdot q^i \;\approx\; e^{-rMp^2(1-p(r-1)/2)}, \tag{9}$$

where the second approximation requires $0 < p \ll 1$ and $q = (1-p)^r$ (cf. section 3.6.3 in [6]). With this we can infer from eq. 6 the second moment of the Willshaw distribution

$$E({X_R}^2) = \sum_{x=0}^{z} x^2 \cdot \sum_{i=0}^{M} p_B(i; M, \frac{k}{n}) \cdot p_B(x; z, 1 - (1 - \frac{k}{n})^i) \tag{10}$$

$$= z^2 - z(2z-1) \cdot (1 - \frac{k^2}{n^2})^M + z(z-1) \cdot \sum_{i=0}^{M} p_B(i; M, \frac{k}{n}) \cdot (1 - \frac{k}{n})^{2i} \tag{11}$$

$$\approx z^2 - z \cdot (2z-1) \cdot (1 - \frac{k^2}{n^2})^M + z \cdot (z-1) \cdot e^{-2M\frac{k^2}{n^2}(1-\frac{k}{2n})} \quad . \tag{12}$$

Finally, for the variance ${\sigma_W}^2$ of the Willshaw distribution the following approximation can be found,

$${\sigma_W}^2 := \mathrm{Var}({X_R}^2) = E({X_R}^2) - E(X_R)^2 \tag{13}$$

$$\approx z p_1(1-p_1) - (z^2 - z)\frac{k}{n}(1-p_1)^2 \ln(1-p_1) \tag{14}$$

$$\approx z p_1(1-p_1) \quad =: {\sigma_B}^2. \tag{15}$$

The first approximation is quite good (cf., Fig. 2b). The second approximation yields the variance ${\sigma_B}^2$ suggested by the classical binomial approximation, and is justified only for small k/n and z (cf. [6]). Thus, the classical analysis of binary associative nets [14,8,11,5] may overestimate the storage capacity and fault tolerance, in particular if the address pattern contains a large number z of one-entries (cf. [6]). On the other hand, this effect could be exploited in a neurophysiological experiment to verify the hypothesis of local cell assemblies in a cortical column. The basic idea would be to stimulate randomly a large neuron number z, and then to compute the coefficient of variation CV:=s.d./mean for the amplitude of the postsynaptic potentials (or currents). If CV$\rightarrow 0$ for increasing z this would falsify the assembly hypothesis, while an asymptotically positive CV would support it.

4 Conclusion and discussion

In order to explain some discrepancies between the true membrane potential distributions and a simple binomial approximation widely used in the literature [14,8,11,5] (for exact approaches see [3,2,12] and the appendix in [8]), a good approximation for the variance of the Willshaw distribution of membrane potentials has been derived (cf. Fig. 2 and eqs. 14 and 15). It turned out that the variance increase is $\sim z^2$ with the stimulation strength z, while the binomial approximation suggests only a linear growth $\sim z$. This result has two important implications (cf. [6]): (1) It improves the classical analysis of binary associative networks which use the binomial approximation and therefore overestimate fault tolerance and storage capacities. (2) It suggests a neurophysiological experiment which could finally prove or disprove the occurrence of local cell assemblies in the cortical areas of the brain. Of course, the analyzed model is very abstract compared to the neurobiological reality (cf. [6]). Whether this approach really leads to viable experiments remains to be seen in future analytical work and discussions with neurophysiologists.

References

[1] V. Braitenberg and A. Schüz. *Anatomy of the cortex. Statistics and geometry.* Springer-Verlag, Berlin, 1991.

[2] J.T. Buckingham. Delicate nets, faint recollections: a study of partially connected associative network memories. *PhD thesis, University of Edinburgh*, 1991.

[3] J.T. Buckingham and D.J. Willshaw. Performance characteristics of associative nets. *Network: Computation in Neural Systems*, 3:407–414, 1992.

[4] D.O. Hebb. *The organization of behavior. A neuropsychological theory.* Wiley, New York, 1949.

[5] A. Knoblauch. Optimal matrix compression yields storage capacity 1 for binary Willshaw associative memory. In O. Kaynak, E. Alpaydin, E. Oja, and L. Xu, editors, *Artificial Neural Networks and Neural Information Processing - ICANN/ICONIP 2003.*, LNCS 2714, pages 325–332. Springer Verlag, Berlin, 2003.

[6] A. Knoblauch. Synchronization and pattern separation in spiking associative memory and visual cortical areas. *PhD thesis, Department of Neural Information Processing, University of Ulm, Germany*, 2003.

[7] A. Knoblauch and G. Palm. Pattern separation and synchronization in spiking associative memories and visual areas. *Neural Networks*, 14:763–780, 2001.

[8] G. Palm. On associative memories. *Biological Cybernetics*, 36:19–31, 1980.

[9] G. Palm. *Neural Assemblies. An Alternative Approach to Artificial Intelligence.* Springer, Berlin, 1982.

[10] G. Palm. Cell assemblies as a guideline for brain research. *Concepts in Neuroscience*, 1:133–148, 1990.

[11] G. Palm. Memory capacities of local rules for synaptic modification. A comparative review. *Concepts in Neuroscience*, 2:97–128, 1991.

[12] F.T. Sommer and G. Palm. Improved bidirectional retrieval of sparse patterns stored by hebbian learning. *Neural Networks*, 12:281–297, 1999.

[13] K. Steinbuch. Die Lernmatrix. *Kybernetik*, 1:36–45, 1961.

[14] D.J. Willshaw, O.P. Buneman, and H.C. Longuet-Higgins. Non-holographic associative memory. *Nature*, 222:960–962, 1969.