Convergence-Zone Episodic Memory: Analysis and Simulations *†

Mark Moll
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA
mmoll+@cs.cmu.edu

Risto Miikkulainen
Department of Computer Sciences
The University of Texas at Austin
Austin, TX 78712 USA
risto@cs.utexas.edu

Abstract

Human episodic memory provides a seemingly unlimited storage for everyday experiences, and a retrieval system that allows us to access the experiences with partial activation of their components. The system is believed to consist of a fast, temporary storage in the hippocampus, and a slow, long-term storage within the neocortex. This paper presents a neural network model of the hippocampal episodic memory inspired by Damasio's idea of Convergence Zones. The model consists of a layer of perceptual feature maps and a binding layer. A perceptual feature pattern is coarse coded in the binding layer, and stored on the weights between layers. A partial activation of the stored features activates the binding pattern, which in turn reactivates the entire stored pattern. For many configurations of the model, a theoretical lower bound for the memory capacity can be derived, and it can be an order of magnitude or higher than the number of all units in the model, and several orders of magnitude higher than the number of binding-layer units. Computational simulations further indicate that the average capacity is an order of magnitude larger than the theoretical lower bound, and making the connectivity between layers sparser causes an even further increase in capacity. Simulations also show that if more descriptive binding patterns are used, the errors tend to be more plausible (patterns are confused with other similar patterns), with a slight cost in capacity. The convergence-zone episodic memory therefore accounts for the immediate storage and associative retrieval capability and large capacity of the hippocampal memory, and shows why the memory encoding areas can be much smaller than the perceptual maps, consist of rather coarse computational units, and be only sparsely connected to the perceptual maps.

1 Introduction

Human memory system can be divided into semantic memory of facts, rules, and general knowledge, and episodic memory that records the individual's day-to-day experiences (Tulving 1972, 1983). Episodic memory is characterized by extreme efficiency and high capacity. New memories are formed every few seconds, and many of those persist for years, even decades (Squire 1987). Another

*An earlier version of this paper appeared in *Neural Networks*, 10:1017–1036, 1997.

†Acknowledgements: We thank Greg Plaxton for pointing us to martingale analysis on this problem, and two anonymous reviewers for references on hippocampal modeling, and for suggesting experiments with sparse connectivity. This research was supported in part by NSF grant #IRI-9309273 and Texas Higher Education Coordinating Board Grant #ARP-444 to the second author. The simulations were run on the Cray Y-MP 8/864 at the University of Texas Center for High-Performance Computing.

significant characteristic of human memory is content-addressability. Most of the memories can be retrieved simply by activating a partial representation of the experience, such as a sound, a smell, or a visual image.

Despite vast amount of research, no clear understanding has yet emerged on exactly where and how the episodic memory traces are represented in the brain. Several recent results, however, suggest that the system consists of two components: the hippocampus serves as a fast, temporary storage where the traces are created immediately as the experiences come in, and the neocortex has the task of organizing and storing the experiences for the lifetime of the individual (Alvarez and Squire 1994; Halgren 1984; Marr 1971; McClelland et al. 1995; Milner 1989; Squire 1992). It seems that the traces are transferred from the hippocampus to the neocortex in a slow and tedious process, which may take several days, or weeks, or even years. After that, the hippocampus is no longer necessary for maintaining these traces, and the resources can be reused for encoding new experiences.

Although several artificial neural network models of associative memory have been proposed (Ackley et al. 1985; Amari 1977, 1988; Anderson 1972; Anderson et al. 1977; Cooper 1973; Grossberg 1983; Gardner-Medwin 1976; Hinton and Anderson 1981; Hopfield 1982, 1984; Kairiss and Miranker 1996; Kanerva 1988; Knapp and Anderson 1984; Kohonen 1971, 1972, 1977, 1989 Kohonen and Mä kisara 1986; Kortge 1990; Little and Shaw 1975; McClelland and Rumelhart 1986b; Miikkulainen 1992; Steinbuch 1961; Willshaw et al. 1969), the fast encoding, reliable associative retrieval, and large capacity of even the hippocampal component of human memory has been difficult to account for. For example in the Hopfield model of N units, $N/4\ln N$ patterns can be stored with a 99% probability of correct retrieval when N is large (Amit 1989; Hertz et al. 1991; Keeler 1988; McEliece et al. 1986). This means that storing and retrieving, for example, 10^6 memories would require in the order of 10^8 nodes and 10^{16} connections, which is unrealistic, given that the hippocampal formation in higher animals such as the rat is estimated to have about 10^6 primary excitatory neurons with 10^{10} connections (Amaral et al. 1990), and the entire human brain is estimated to have about 10^{11} neurons and 10^{15} synapses (Jessell 1991).

These earlier models had a uniform, abstract structure and were not specifically motivated by any particular part of the human memory system. In this paper, a new model for associative episodic memory is proposed that makes use of three ideas about how the hippocampal memory might be put together. The model abstracts most of the low-level biological circuitry, focusing on showing that with a biologically motivated overall architecture, an episodic memory model exhibits capacity and behavior very similar to that of the hippocampal memory system. The three central ideas are: (1) value-unit encoding in the input feature maps, (2) sparse, random encoding of traces in the hippocampus, and (3) a convergence-zone structure between them.

Since the input to the memory consists of sensory experience, in the model it should have a representation similar to the perceptual representations in the brain. The low-level sensory representations are organized into maps, that is, similar sensory inputs are represented by nearby locations on the cortical surface (Knudsen et al. 1987). It is possible that also higher-level representations have a map-like structure. This is hard to verify, but at least there is plenty of support for value-unit encoding, that is, that the neurons respond selectively to only certain types of inputs, such as particular faces, or facial expressions, or particular words (Hasselmo et al. 1989; Heit et al. 1989; Rolls 1984).

The structure of the hippocampus is quite well known, and recently its dynamics in memory processing have also been observed. Wilson and McNaughton (1993) found that rats encode locations in the maze through ensembles of seemingly random, sparse activation patterns in the hippocampal area CA1. When the rat explores new locations, new activation patterns appear, and when it returns to the earlier locations, the same pattern is activated as during the first visit. O'Reilly and McClelland (1994) showed that the hippocampal circuitry is well-designed to form such sparse, diverse encodings, and that it can also perform pattern completion during recall.

Damasio (1989b, 1989a) proposed a general framework for episodic representations, based on observations of typical patterns of injury-related deficits. The idea is that there is no multi-modal cortical area that would build an integrated and independent representation of an experience from its low-level sensory representations. Instead, the representation takes place only in the low-level cortices, with the different parts bound together by a hierarchy of convergence zones. An episodic representation can be recreated by activating its correspoding binding pattern in the convergence zone.

The convergence-zone episodic memory model is loosely based on the above three ideas. It consists of a layer of perceptual maps and a binding layer. An episodic experience appears as a pattern of local activations across the perceptual maps, and is encoded as a sparse, random pattern in the binding layer. The connections between the maps and the binding layer store the encoding in a single presentation, and the complete perceptual pattern can later be regenerated from partial activation of the input layer.

Many details of the low-level neural circuitry are abstracted in the model. The units in the model correspond to functional columns rather than neurons and their activation levels are represented by integers. Multi-stage connections from the perceptual maps to the hippocampus are modeled by direct binary connections that are bidirectional, and the connections within the hippocampus are not taken into account. At this level of abstraction, the behavior of the model can be analyzed both theoretically and experimentally, and general results can be derived about its properties.

A theoretical analysis shows that: (1) with realistic-size maps and binding layer, the capacity of the convergence-zone memory can be very high, higher than the number of units in the model, and can be several orders of magnitude higher than the number of binding-layer units, (2) the majority of the neural hardware is required in the perceptual processing; the binding layer needs to be only a fraction of the size of the perceptual maps, and (3) the computational units could be very coarse in the hippocampus and in the perceptual maps; the required capacity is achieved with a very small number of such units. Computational simulations of the model further suggest that (1) the average storage capacity may be an order of magnitude higher than the theoretical lower bound, (2) the capacity can be further increased by reducing the connectivity between feature maps and the binding layer, with best results when the connectivity matches the sparseness of the binding representations; and (3) if the binding patterns for similar inputs are made more similar, the errors that the model makes become more plausible: the retrieved patterns are similar to the correct patterns. These results suggest how one-shot storage, content-addressability, high capacity, and robustness could all be achieved within the resources of the hippocampal memory system.

2 Outline of the Model

The convergence-zone memory model consists of two layers of real-valued units (the feature map layer and the binding layer), and bidirectional binary connections between the layers (figure 1). Perceptual experiences are represented as vectors of feature values, such as color=`red`, shape=`round`, size=`small`. The values are encoded as units on the feature maps. There is a separate map for each feature domain, and each unit on the map represents a particular value for that feature. For

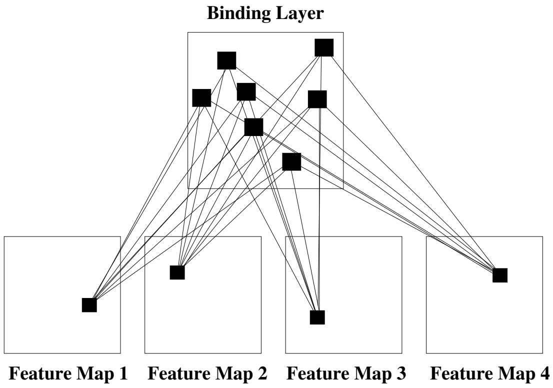


Figure 1: **Storage.** The weights on the connections between the appropriate feature units and the binding representation of the pattern are set to 1.

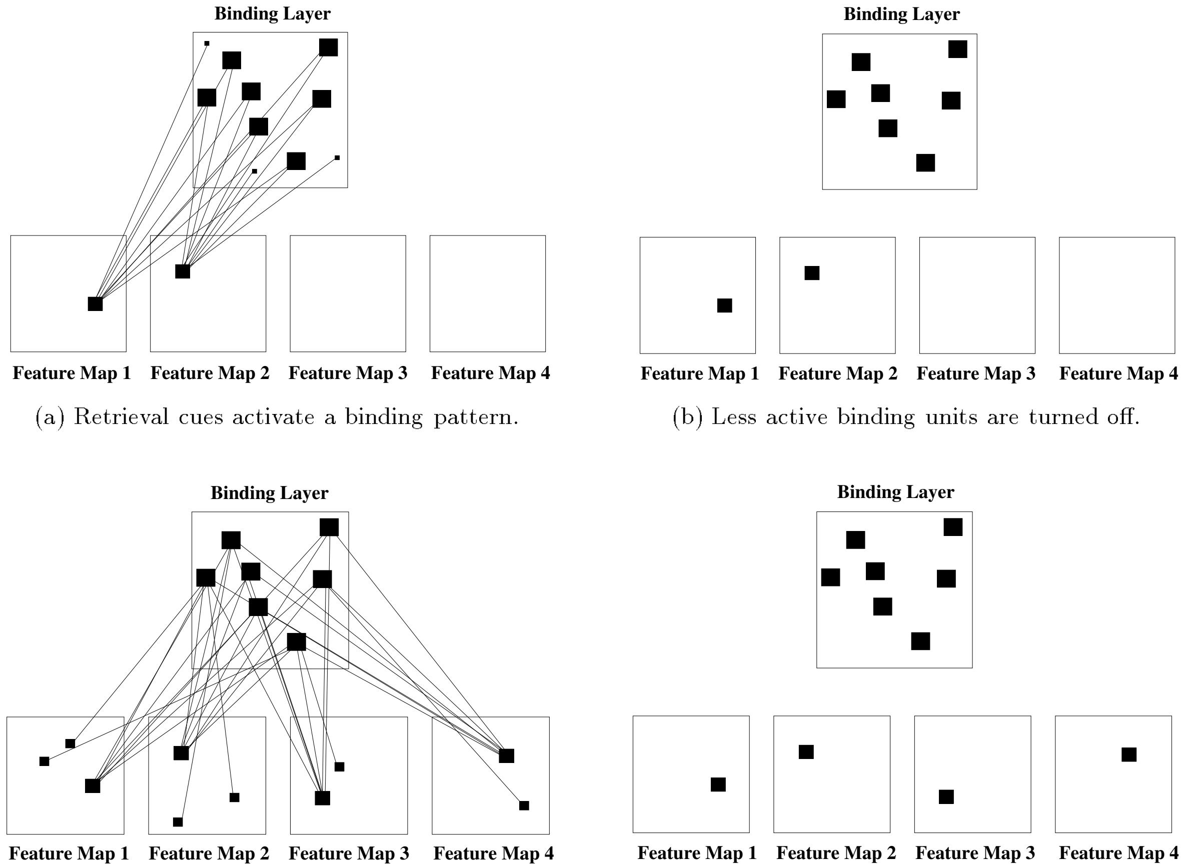
instance, on the map for the color feature, the value `red` could be specified by turning on the unit in the lower-right quarter (figure 1). The feature map units are connected to the binding layer with bidirectional binary connections (i.e. the weight is either 0 or 1). An activation of units in the feature map layer causes a number of units to become active in the binding layer, and vice versa. In effect, the binding layer activation is a compressed, distributed encoding of the perceptual value-unit representation.

Initially, all connections are inactive at 0. A perceptual experience is stored in the memory through the feature map layer in three steps. First, those units that represent the appropriate feature values are activated at 1. Second, a subset of m binding units are randomly selected in the binding layer as the compressed encoding for the pattern, and activated at 1. Third, the weights of all the connections between the active units in the feature maps and the active units in the binding layer are set to 1 (figure 1). Note that only one presentation is necessary to store a pattern this way.

To retrieve a pattern, first all binding units are set to 0. The pattern to be retrieved is partially specified in the feature maps by activating a subset of its feature units. For example, in figure 2a the memory is cued with the two leftmost features. The activation propagates to the binding layer through all connections that have been turned on so far. The set of binding units that a particular feature unit turns on is called the *binding constellation* of that unit. All binding units in the binding encoding of the pattern to be retrieved are active at 2 because they belong to the binding constellation of both retrieval cue units. A number of other units are also activated at 1, because each cue unit takes part in representing multiple patterns, and therefore has several other active connections as well. Only those units active at 2 are retained; units with less activation are turned off (figure 2b).

The activation of the remaining binding units is then propagated back to the feature maps (figure 2c). A number of units are activated at various levels in each feature map, depending on how well their binding constellation matches the current pattern in the binding layer. Chances are that the unit that belongs to the same pattern as the cues has the largest overlap and becomes most highly activated. Only the most active unit in each feature map is retained, and as a result, a complete, unambiguous perceptual pattern is retrieved from the system (figure 2d).

If there are n units in the binding layer and m units are chosen as a representation for a pattern, the number of possible different binding representations is equal to $\binom{n}{m}$. If n is sufficiently



(a) Retrieval cues activate a binding pattern.

(b) Less active binding units are turned off.

(c) Binding pattern activates feature units.

(d) Less active feature units are turned off.

Figure 2: **Retrieval.** A stored pattern is retrieved by presenting a partial representation as a cue. The size of the square indicates activation level of the unit.

large and m is relatively small compared to n, this number is extremely large, suggesting that the convergence-zone memory could have a very large capacity.

However, due to the probabilistic nature of the storage and retrieval processes, there is always a chance that the retrieval will fail. The binding constellations of the retrieval cue units may overlap significantly, and several spurious units may be turned on at the binding layer. When the activation is propagated back to the feature maps, some random unit in a feature map may have a binding constellation that matches the spurious units very well (figure 3). This *rogue* unit may receive more activation than the correct unit, and a wrong feature value may be retrieved. As more patterns are stored, the binding constellations of the feature units become larger, and erroneous retrieval becomes more likely.

To determine the capacity of the convergence-zone memory, the chance of retrieval error must be computed. Below, a probabilistic formulation of the model is first given, and a lower bound for the retrieval error is derived.

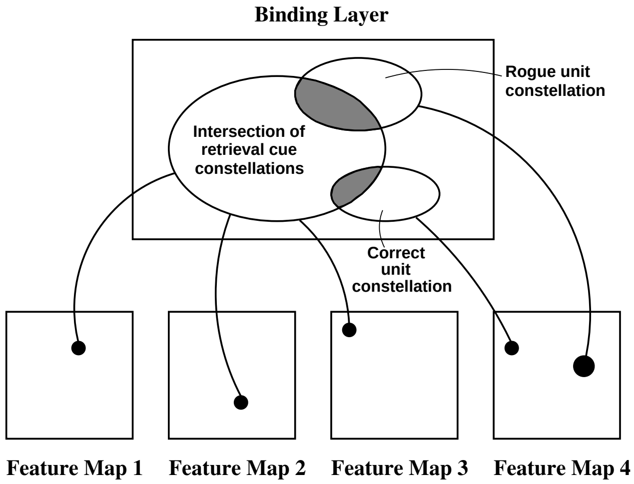


Figure 3: **Erroneous retrieval.** A rogue feature unit is retrieved, instead of the correct one, when its binding constellation has a more units in common with the intersection of the retrieval cue constellations than the binding constellation of the correct unit.

3 Probabilistic Formulation

Let Z_i be the size of the binding constellation of a feature unit after i patterns have been stored on it and let Y_i be its increase after storing the ith pattern on it. Obviously, $Y_1 = m$. To obtain the distribution of Y_i when $i > 1$, note that the new active connections belong to the intersection of a randomly chosen subset of m connections among all n connections of the unit, and its remaining inactive connections (a set with $n - z_{i-1}$ elements, where z_{i-1} is the binding constellation at the previous step). Therefore, $Y_i, i > 1$ is hypergeometrically distributed (appendix A.1) with parameters $m, n - z_{i-1}$, and n:

$$\mathrm{P}(Y_i = y | Z_{i-1} = z_{i-1}) = \binom{n - z_{i-1}}{y} \binom{z_{i-1}}{m-y} \Big/ \binom{n}{m}. \tag{1}$$

The constellation size Z_i is then given by

$$Z_i = \sum_{k=1}^{i} Y_k. \tag{2}$$

Let I be the number of patterns stored on a particular feature unit after p random feature patterns have been stored in the entire memory. I is binomially distributed (appendix A.1) with parameters p and $\frac{1}{f}$, where f is the number of units in a feature map:

$$\mathrm{P}(I = i) = \binom{p}{i} \left(\frac{1}{f}\right)^{i} \left(1 - \frac{1}{f}\right)^{(p-i)}. \tag{3}$$

Let Z be the binding constellation of a particular feature unit after p patterns have been stored in the memory. It can be shown (appendix A.2) that

$$\begin{aligned} \mathrm{E}(Z) &= n(1 - (1 - \frac{m}{nf})^p) \quad \text{and} && (4)\\ \mathrm{Var}(Z) &= n(1 - \tfrac{m}{nf})^p(1 - n(1 - \tfrac{m}{nf})^p) + n(n-1)\left(1 - \tfrac{m(2n-m-1)}{n(n-1)f}\right)^p, && (5) \end{aligned}$$

Initially, when no patterns are stored, the binding constellation is zero and it will converge to n as more patterns are stored (since $0 < 1 - \frac{m}{nf} < 1$). Because the bases of the exponentials in the variance of Z are smaller than 1, the variance will go to zero when p goes to infinity. Therefore, in the limit the binding constellation will cover the entire binding layer with probability 1.

The binding constellation of a feature unit, given that at least one pattern has been stored on it, is denoted by $\tilde{Z}$. This variable represents the binding constellation of a retrieval cue, which necessarily must have at least one pattern stored on it (assuming that the retrieval cues are valid). The expected value and variance of $\tilde{Z}$ follow from equations 4 and 5:

$$\begin{aligned} \mathrm{E}(\tilde{Z}) &= m + (n-m)(1-(1-\frac{m}{nf})^p) \quad \text{and} && (6)\\ \mathrm{Var}(\tilde{Z}) &= (n-m)(1-\tfrac{m}{nf})^{p-1}(1-(n-m)(1-\tfrac{m}{nf})^{p-1}) \\ &\quad +(n-m)(n-m-1)\Big(1-\tfrac{m(2n-m-1)}{n(n-1)f}\Big)^{p-1}. && (7) \end{aligned}$$

Note that the expected value of $\tilde{Z}$ is always larger than that of Z. Initially the difference is exactly m, and it goes to zero as p goes to infinity (because $\tilde{Z}$ also converges to n).

Let $\tilde{Z}^j$ be the binding constellation of the jth retrieval cue and let X_j be the number of units in the intersection of the first j retrieval cues. Clearly, $X_1 = \tilde{Z}^1 \geq m$. To get X_j for $j > 1$, we remove from consideration the m units all retrieval cues necessarily have in common (because they belong to the same stored pattern), and randomly select $\tilde{z}^j - m$ units from the remaining set of $n - m$ units and see how many of them belong to the current intersection of $x_{j-1} - m$ units. This is a hypergeometric distribution with parameters $\tilde{z}^j - m, x_{j-1} - m$, and $n - m$:

$$\mathrm{P}(X_j = x_j | \tilde{Z}^j = \tilde{z}^j, X_{j-1} = x_{j-1}) = \binom{x_{j-1}-m}{x_j - m}\binom{n - x_{j-1}}{\tilde{z}^j - x_j} \Big/ \binom{n-m}{\tilde{z}^j - m}. \qquad (8)$$

The size of the total binding constellation activated during retrieval is obtained by taking this intersection over the binding constellations of all c retrieval cues.

The number of units in common between a potential rogue unit and the c retrieval cues is denoted by R_{c+1} and is also hypergeometrically distributed, however with parameters z, x_c, and n, because we cannot assume that the rogue unit has at least m units in common with the cues:

$$\mathrm{P}(R_{c+1} = r | Z = z, X_c = x_c) = \binom{x_c}{r}\binom{n - x_c}{z - r} \Big/ \binom{n}{z}. \qquad (9)$$

The correct unit in a retrieval map (i.e. in a feature map where a retrieval cue was not presented and where a feature value needs to be retrieved) will receive an activation X_{c+1}, because it also has at least m units in common with the retrieval cues. The correct unit will be retrieved if $X_{c+1} > R_{c+1}$. Now, X_{c+1} and R_{c+1} differ only in the last intersection step, where X_{c+1} depends on $\tilde{Z}$ and X_c, and R_{c+1} depends on Z and X_c. Since $\mathrm{E}(\tilde{Z}) > \mathrm{E}(Z)$ (equations 4 and 6), $\mathrm{E}(X_{c+1}) > \mathrm{E}(R_{c+1})$, and the correct unit will be retrieved most of the time, although this advantage gradually decreases as more patterns are stored in the memory. In each feature map there are $(f-1)$ potential rogue units, so the conditional probability of successful retrieval is $(1 - \mathrm{P}(R_{c+1} > X_{c+1} | X_{c+1}, Z, X_c))^{(f-1)}$, not addressing tie-breaking. Unfortunately, it is very difficult to compute p_{success}, the unconditional probability of successful retrieval, because the distribution functions of Z, X_c, X_{c+1} and R_{c+1} are not known. However, it is possible to derive bounds for p_{success} and show that with reasonable values for n, m, f, and p, the memory is reliable.

4 Lower Bound for Memory Capacity

Memory capacity can be defined as the maximum number of patterns that can be stored in the memory so that the probability of correct retrieval with a given number of retrieval cues is greater than α (a constant close to 1). In this section, a lower bound for the chance of successful retrieval will be derived. The analysis consists of three steps: (1) bounds for the number of patterns stored on a feature unit; (2) bounds for the binding constellation size; and (3) bounds for the intersections of binding constellations. Given particular values for the system parameters, and ignoring dependencies among constellations, it is then possible to derive a lower bound for the capacity of the model.

4.1 Number of patterns stored on a feature unit

Since I is binomially distributed (with parameters p and $\frac{1}{f}$), Chernoff bounds (appendix A.1) can be applied:

$$\mathrm{P}(I \leq (1-\delta_1)\frac{p}{f}) \leq \left[\frac{e^{-\delta_1}}{(1-\delta_1)^{1-\delta_1}}\right]^{\frac{p}{f}}, \qquad 0 < \delta_1 < 1, \tag{10}$$

$$\mathrm{P}(I \geq (1+\delta_2)\frac{p}{f}) \leq \left[\frac{e^{\delta_2}}{(1+\delta_2)^{1+\delta_2}}\right]^{\frac{p}{f}}, \qquad \delta_2 > 0. \tag{11}$$

These equations give the probability that I is more than $\delta_1\frac{p}{f}$ and $\delta_2\frac{p}{f}$ off its mean. The parameters δ_1 and δ_2 determine the tradeoff between the tightness of the bounds and the probability of satisfying them. If bounds are desired with a given probability β, the right hand sides of equations 10 and 11 are made equal to β and solved for δ_1 and δ_2. The lower and upper bound for the number of patterns stored on a feature unit then are

$$i_l = \begin{cases} (1-\delta_1)\frac{p}{f} & \text{if a solution for } \delta_1 \text{ exists} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

$$i_u = (1+\delta_2)\frac{p}{f}. \tag{13}$$

Given that at least one pattern has been stored on a feature unit the bounds become

$$\tilde{i}_l = \begin{cases} 1+(1-\delta_1)\frac{p-1}{f} & \text{if a solution for } \delta_1 \text{ exists} \\ 1 & \text{otherwise} \end{cases} \tag{14}$$

$$\tilde{i}_u = 1+(1+\delta_2)\frac{p-1}{f} \tag{15}$$

4.2 Size of the binding constellation

Instead of choosing exactly m different binding units for the binding constellation of a feature map unit, consider the process of randomly selecting k not-necessarily-distinct units in such a way that the expected number of different units is m. This will make the analysis easier at the cost of larger variance, but the bounds derived will also be valid for the actual process. To determine k, note that the number of units that do not belong to the binding representation is equal to $n-m$ on average:

$$n(1-\frac{1}{n})^k = n-m. \tag{16}$$

Solving for k, we get

$$k = \frac{\ln n - \ln(n-m)}{\ln n - \ln(n-1)}. \tag{17}$$

Note that k is almost equal to m for large n.

Let us assume i patterns are stored on the feature map unit, which is equivalent to selecting ki units from the binding layer at random. Let Z^{E}_v be the expected size of the final binding constellation, estimated after v binding units have been selected. Then

$$\begin{aligned} Z^{\mathrm{E}}_v &= Z'_v + (n - Z'_v)(1 - (1 - \frac{1}{n})^{ki-v}) \\ &= n - (n - Z'_v)(1 - \frac{1}{n})^{ki-v}, \end{aligned} \tag{18}$$

where Z'_v is the size of the binding constellation formed by the first v selected units. Obviously Z'_v is equal to Z'_{v-1} or exactly one larger, and the expected increase of Z'_v is $1 - \frac{Z'_{v-1}}{n}$. Since Z^{E}_{v-1} depends stochastically only on Z'_{v-1}, the expected value of Z^{E}_v, given Z^{E}_{v-1}, is

$$\begin{aligned} \mathrm{E}(Z^{\mathrm{E}}_v | Z^{\mathrm{E}}_{v-1} = z^{\mathrm{E}}_{v-1}) &= \mathrm{E}(Z^{\mathrm{E}}_v | Z'_{v-1} = z'_{v-1}) \\ &= n - \left(n - z'_{v-1} - (1 - \frac{z'_{v-1}}{n})\right)(1 - \frac{1}{n})^{ki-v} \\ &= n - (n - z'_{v-1})(1 - \frac{1}{n})^{ki-v+1} \\ &= z^{\mathrm{E}}_{v-1}. \end{aligned} \tag{19}$$

Therefore, $\mathrm{E}(Z^{\mathrm{E}}_v | Z^{\mathrm{E}}_{v-1}) = Z^{\mathrm{E}}_{v-1}$ and the sequence of variables $Z^{\mathrm{E}}_0, \ldots, Z^{\mathrm{E}}_{ki}$ is a martingale (see appendix A.3). Moreover, it can be shown (appendix A.4) that $|Z^{\mathrm{E}}_v - Z^{\mathrm{E}}_{v-1}| \leq 1$, so that bounds for the final binding constellation Z can be obtained from Azuma's inequalities. For the lower bound, the martingale $Z^{\mathrm{E}}_0, \ldots, Z^{\mathrm{E}}_{ki_l}$ (with length ki_l) is used, and for the upper bound, $Z^{\mathrm{E}}_0, \ldots, Z^{\mathrm{E}}_{ki_u}$ (with length ki_u). Using equation 18 and noting that $Z = Z^{\mathrm{E}}_{ki_l}$ for the lower bound and $Z = Z^{\mathrm{E}}_{ki_u}$ for the upper bound, Azuma's inequalities can be written as:

$$\begin{aligned} \mathrm{P}(Z^{\mathrm{E}}_{ki_l} \leq Z^{\mathrm{E}}_0 - \lambda\sqrt{ki_l}) &= \mathrm{P}(Z \leq n(1 - (1 - \frac{1}{n})^{ki_l}) - \lambda\sqrt{ki_l}) \\ &\leq e^{-\lambda^2/2}, \qquad \lambda > 0, \end{aligned} \tag{20}$$

$$\begin{aligned} \mathrm{P}(Z^{\mathrm{E}}_{ki_u} \geq Z^{\mathrm{E}}_0 + \lambda\sqrt{ki_u}) &= \mathrm{P}(Z \geq n(1 - (1 - \frac{1}{n})^{ki_u}) + \lambda\sqrt{ki_u}) \\ &\leq e^{-\lambda^2/2}, \qquad \lambda > 0. \end{aligned} \tag{21}$$

After deriving a value for λ based on the desired confidence level β, the following lower and upper bounds for Z are obtained:

$$z_l = n(1 - (1 - \frac{1}{n})^{ki_l}) - \lambda\sqrt{ki_l} \tag{22}$$

$$z_u = n(1 - (1 - \frac{1}{n})^{ki_u}) + \lambda\sqrt{ki_u}. \tag{23}$$

The corresponding bounds for $\tilde{Z}$ are:

$$\tilde{z}_l = n(1 - (1 - \frac{1}{n})^{k\tilde{i}_l}) - \lambda\sqrt{k\tilde{i}_l} \tag{24}$$

$$\tilde{z}_u = n(1 - (1 - \frac{1}{n})^{k\tilde{i}_u}) + \lambda\sqrt{k\tilde{i}_u}. \tag{25}$$

4.3 Intersection of binding constellations

The process of forming the intersection of c binding constellations incrementally one cue at a time can also be formulated as a martingale process. To see how, consider the process of forming an intersection of two subsets of a common superset incrementally, by checking (one at a time) whether each element of the first set occurs in the second set. Assume that v elements have been checked this way. Let X'_v denote the number of elements found to be in the intersection so far, and X^{E}_v the currently expected number of elements in the final intersection. Then

$$X^{\mathrm{E}}_v = X'_v + \frac{(n_1 - v)(n_2 - X'_v)}{n_s - v}, \tag{26}$$

where n_1, n_2 and n_s are the sizes of the first, second, and the superset. As shown in appendix A.5, the sequence $X^{\mathrm{E}}_0, \ldots, X^{\mathrm{E}}_{n_1}$ is a martingale. In addition, if $n_1 + n_2 - 1 < n_s$, $|X^{\mathrm{E}}_v - X^{\mathrm{E}}_{v-1}| \leq 1$, and Azuma's inequalities can be applied.

The above process applies to forming the intersection of binding constellations of retrieval cues when the intersection in the previous step is chosen as the first set, the binding constellation of the jth cue as the second set, the binding layer as the common superset, and the m units all retrieval cues have in common are excluded from the intersection. In this case,

$$\begin{aligned} n_1 &= x_{j-1,u} - m & (27)\\ n_2 &= \tilde{z}_u - m & (28)\\ n_s &= n - m, & (29) \end{aligned}$$

where $x_{j-1,u}$ is an upper bound for X_{j-1}. Azuma's inequality can be applied if $x_{j-1,u} + \tilde{z}_u - 1 < n$ (which needs to be checked). Using equations 27-29 in 26 and noting that $X^{\mathrm{E}}_{n_1} = X_j - m$, Azuma's inequality becomes

$$\begin{aligned} \mathrm{P}(X^{\mathrm{E}}_{n_1} \geq X^{\mathrm{E}}_0 - \lambda\sqrt{n_1}) &= \mathrm{P}(X_j \geq m + \frac{(x_{j-1,u} - m)(\tilde{z}_u - m)}{(n-m)} + \lambda\sqrt{x_{j-1,u} - m}) \\ &\leq e^{-\lambda^2/2}, \qquad \lambda > 0. \end{aligned} \tag{30}$$

After deriving value for λ based on the desired confidence, the following upper bound for X_j is obtained:

$$x_{j,u} = m + \frac{(x_{j-1,u} - m)(\tilde{z}_u - m)}{(n-m)} + \lambda\sqrt{x_{j-1,u} - m} \tag{31}$$

This bound is computed recursively, with $x_{1,u} = m$. Comparing with the probabilistic formulation of section 3, note that $\frac{(x_{j-1,u}-m)(\tilde{z}_u-m)}{(n-m)}$ is the expected value of the hypergeometric distribution derived for X_j (equation 8) when $\tilde{Z}_j$ and X_{j-1} are at their upper bounds.

As the last step in the analysis of binding constellations, the bounds for X_{c+1} and R_{c+1} must be computed. When X_c is at its upper bound, the intersection is the largest, and a potential rogue unit has the largest chance of taking over. In this case, a lower bound for X_{c+1} is obtained by carrying the intersection process one step further, and applying Azuma's inequality:

$$\mathrm{P}(X_{c+1} \leq m + \frac{(x_{c,u} - m)(\tilde{z}_l - m)}{(n-m)} - \lambda\sqrt{x_{c,u} - m}) \leq e^{-\lambda^2/2}, \qquad \lambda > 0. \tag{32}$$

which results in

$$x_{c+1,l} = m + \frac{(x_{c,u} - m)(\tilde{z}_l - m)}{(n-m)} - \lambda\sqrt{x_{c,u} - m} \tag{33}$$

If the resulting lower bound is smaller than m, m can be used instead. Similarly, to get the upper bound for R_{c+1}, one more intersection step needs to be carried out, but this time the m units are not excluded:

$$\mathrm{P}(R_{c+1} \geq \frac{x_{c,u} z_u}{n} + \lambda\sqrt{x_{c,u}}) \leq e^{-\lambda^2/2}, \qquad \lambda > 0, \tag{34}$$

and the upper bound becomes

$$r_{c+1,u} = \frac{x_{c,u} z_u}{n} + \lambda\sqrt{x_{c,u}}. \tag{35}$$

4.4 Dependencies between binding constellations

Strictly speaking, the above analysis is valid only when the binding constellations of each cue are independent. If the same partial pattern is stored multiple times, the constellations will overlap beyond the m units that they necessarily have in common. Such overlap tends to increase the size of the final intersection.

In most cases of realistic size, however, the increase is negligible. The number of features V in common between two random patterns of c features each is given by the binomial distribution:

$$\mathrm{P}(V = v) = \binom{c}{v} \left(\frac{1}{f}\right)^v \left(1 - \frac{1}{f}\right)^{c-v}. \tag{36}$$

The chance that two random patterns of c features have more than one feature in common is

$$\begin{aligned} \mathrm{P}(V > 1) &= 1 - \mathrm{P}(V = 0) - \mathrm{P}(V = 1) \\ &= 1 - \left(1 - \frac{1}{f}\right)^c - c\left(\frac{1}{f}\right)\left(1 - \frac{1}{f}\right)^{c-1}, \end{aligned} \tag{37}$$

which can be rewritten as

$$\mathrm{P}(V > 1) = 1 - (1 + \frac{c}{f-1})(1 - \frac{1}{f})^c. \tag{38}$$

This chance is negligible for sufficiently large values of f. For example, already when $f = 5,000$ and $c = 3$, the chance is 1.2×10^{-7}, and can be safely ignored when computing a lower bound for the capacity.

4.5 Obtaining the lower bound

It is now possible to use equations 10–15, 17 and 20–25, and 30–35 to derive a lower bound for the probability of successful retrieval with given system parameters n, m, f, t, c, and p, where t is the total number of feature maps. The retrieval is successful if $r_{c+1,u}$, the upper bound for R_{c+1}, is lower than $x_{c+1,u}$, the lower bound for X_{c+1}. Under this constraint, the probability that none of the variables in the analysis exceeds its bounds is a lower bound for successful retrieval.

Obtaining the upper bound for X_c involves bounding $3c-1$ variables: I and $\tilde{Z}$ for the c cues and X_c for the $c-1$ intersections. Computing $x_{c+1,l}$ and $r_{c+1,u}$ each involve bounding 3 variables (I, Z, and X_{c+1}; I, $\tilde{Z}$, and R_{c+1}). There are $t-c$ maps, each with one $x_{c+1,l}$ bound and $f-1$ different $r_{c+1,u}$ bounds (one for each rogue unit). The total number of bounds is therefore $3c - 1 + 3f(t-c)$. Setting the righthand sides of the inequalities 10–11, 20–21, 30, 32, and 34 equal to a small constant β, a lower bound for successful retrieval is obtained:

$$p_{\text{success}} > (1 - \beta)^{3c-1+3f(t-c)}, \tag{39}$$

which, for small β, can be approximated by

$$p_{\text{success}} > 1 - (3c - 1 + 3f(t - c))\beta. \tag{40}$$

On the other hand, if it is necessary to determine a lower bound for the capacity of a model with given n, m, f, t, and c at a given confidence level p_{success}, β is first obtained from equation 40, and the number of patterns p is then increased until one of the bounds 10–11, 20–21, 30, 32, or 34 is exceeded, or $r_{c+1,u}$ becomes greater than $x_{c+1,l}$.

5 Example: Modeling the Hippocampal Memory System

As an example, let us apply the above analysis to the hippocampal memory system. It is difficult to estimate how coarse the representations are in such a system, and how many effective computational units and connections there should be. The numbers of neurons and connections in the rat hippocampal formation have been used as a guideline below. Although the human hippocampus is certainly larger than that of the rat, the hippocampus, being phylogenetically one the oldest areas of the brain, is fairly similar across higher mammals and should give an indication of the orders of magnitude involved. More importantly, the convergence-zone model can be shown to apply to a wide range of these parameters. Two cases at the opposite ends of the spectrum are analyzed below: one where the number of computational units and connections is assumed to be limited, and another that is based on a large number of effective units and connections.

5.1 A Coarse-Grained Model

First note that each unit in the model is meant to correspond to a vertical column in the cortex. It is reasonable to assume feature maps with 10^6 of such columns (Sejnowski and Churchland 1989). Each input activates a local area on the map, including perhaps 10^2 columns above threshold. Therefore, the feature maps could be approximated with 10^4 computational units. There would be a minimum of perhaps 4 such maps, of which 3 could be used to cue the memory.

There are some 10^6 primary excitatory cells in the rat hippocampal formation (Amaral et al. 1990,Boss et al. 1985, 1987;Squire et al. 1989). If we assume that functional units contain 10^2 of them, then the model should have 10^4 binding units. Only about 0.5-2.5% of the hippocampal neurons are simultaneously highly active (O'Reilly and McClelland 1994), so a binding pattern of 10^2 units would be appropriate. Assuming that all computational units in the feature maps are connected to all units in the hippocampus, there are a total of 10^8 afferent connections to the hippocampus, and the number of such connections per vertical column in the feature maps and per excitatory neuron in the hippocampus is 10^2, both of which are small but possible numbers (Amaral et al. 1990).

If we select $f = 17,000$, $n = 11,500$, $m = 150$, and store 1.5×10^4 patterns in the memory, $\tilde{z}_u$ and $x_{j-1,u}$ are less than $\frac{1}{2}n$, the chance of partial overlap of more than 1 feature is less than 1.04×10^{-8}, and the analysis above is valid. Setting $\beta = 1.96 \times 10^{-7}$ yields bounds $r_{c+1,u} < x_{c+1,l}$ with $p_{\text{success}} > 99\%$. In other words, 1.5×10^4 traces can be stored in the memory with 99% probability of successful retrieval. Such a capacity is approximately equivalent of storing one new memory every 15 seconds for 4 days, 16 hours a day, which is similar to what is required from the human hippocampal system.

5.2 A Fine-Grained Model

It is possible that a lot more neurons and connections are involved in the hippocampal memory system than assumed above. For example, let us assume that each of the vertical columns in the feature maps is computationally distinctive, that is, there are 10^6 units in the feature maps. Let us further assume that the system has 15 feature maps, 10 of which is used to cue the memory, and the binding layer consists of 10^5 units, with 150 used for each binding pattern. Assuming full connectivity between the feature units and the binding units, there are 1.5×10^{12} connections in the system, which might be possible if it is assumed that a large number of collaterals exist on the inputs to the hippocampus.

Applying the above analysis to this memory configuration, 0.85×10^8 patterns can be stored with 99% probability of successful retrieval.[1] In other words, a new trace could be stored every 15 seconds for 62 years, 16 hours a day, without much memory interference.

This kind of capacity is probably enough for the entire human lifetime, and exceeds the requirements for the hippocampal formation. With such a capacity, there would be no need to transfer representations to the neocortical memory system. One conclusion from this analysis is that the hippocampal formation is likely to have a more coarse-grained than fine-grained structure. Another conclusion is that it is possible that the neocortical memory component may also be based on convergence zones. The result is interesting also from the theoretical point of view, because the lower bound is an order of magnitude higher than the number of units in the system, and three orders of magnitude higher than the number of binding units. To our knowledge, this lower bound is already higher than what is practically possible with other neural network models of associative memory to date.

6 Experimental Average Capacity

The analysis above gives us a lower bound for the capacity of the convergence-zone memory; the average-case capacity may be much higher. Although it is difficult to derive the average capacity theoretically, an estimate can be obtained through computer simulation. Not all configurations of the model can be simulated, though. The model has to be small enough to fit in the available memory, while at the same time fulfilling the assumptions of the analysis so that lower bounds can be obtained for the same model.

To find such configurations, first the feature map parameters f, t, and c, and the confidence level β are fixed to values such that $p_{\text{success}} = 99\%$ (equation 40). Second, a value for n is chosen so that the model will fit in the available memory. The connections take up most of the memory space (even if each connection is represented by one bit) and the amount of memory allocated for the feature map and the binding layer activations, the array of patterns, and the simulation program itself is negligible. Finally, the size of the binding pattern m and the maximum number of patterns p is found such that the theoretical bounds yield $r_{c+1,u} < x_{c+1,l}$ and the partial overlap is negligible. In the models studied so far, the highest capacity has been obtained when m is only be a few percent of the size of the binding layer, as in the hippocampus.

The simulation program is straightforward. The activations in each map are represented as arrays of integers. The connections between a feature map and the binding layer are encoded

[1]In this case, $\tilde{z}_u$ and $x_{j-1,u}$ are less than $\frac{1}{2}n$, the chance of partial overlap of more than 1 feature is less than 0.45×10^{-10}, and setting $\beta = 0.5 \times 10^{-9}$ yields bounds $r_{c+1,u} < x_{c+1,l}$ with $p_{success} > 99\%$.

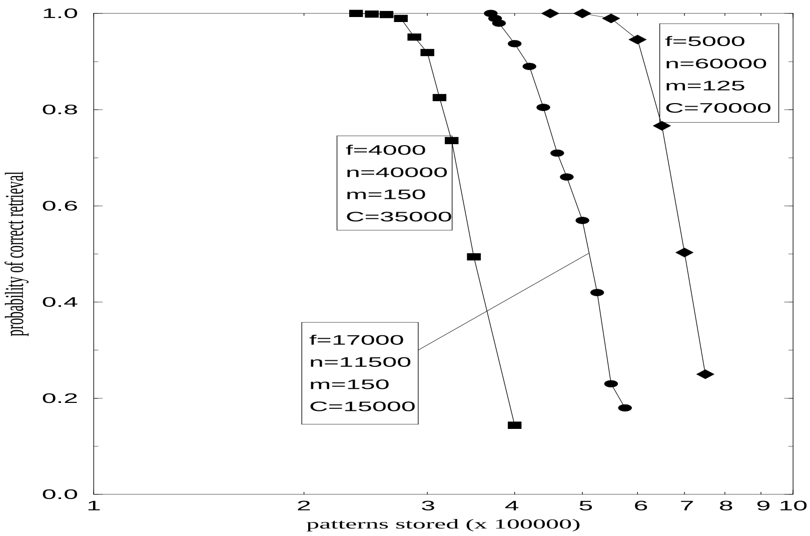


Figure 4: **Experimental average capacity.** The horizontal axis shows the number of patterns stored in a logarithmic scale of hundreds of thousands. The vertical axis indicates the percentage of correctly retrieved patterns out of a randomly-selected subset of 500 stored patterns (a different set was selected each time). Each model consisted of 4 feature maps, and during retrieval, the fourth feature was retrieved using the first 3 as cues. The models differed in the sizes of the feature maps f, binding layer size n, and binding pattern size m. The plots indicate averages over 3 simulations. The theoretical capacity C of the first model with $p_{\text{success}} = 99\%$ was 35,000, that of the second $15,000$, and that of the third $70,000$.

as a two-dimensional array of bits, one bit for each connection. Before the simulation, a set of $p_{\max}$ random patterns are generated as a two-dimensional array of $p_{\max} \times t$ integers. A simulation consists of storing the patterns one at a time and periodically testing how many of a randomly-selected subset of them can be correctly retrieved with a partial cue.

The "fine-grained" example of section 5.2 is unfortunately too large to simulate. With $15 \times 10^5 \times 10^6 = 1.5 \times 10^{12}$ connections it would require 187.5 Gigabytes of memory, which is not possible with the current computers. However, the "coarse-grained" model has $4 \times 17,000 \times 11,500 = 7.82 \times 10^8$ one-bit connections, which amounts to approximately 100MB, and easily fits in available memory.

Several configurations were simulated, and they all gave qualitatively similar results (figure 4). In the coarse-grained model, practically no retrieval errors were produced until 370,000 patterns had been stored. With 375,000 patterns, 99% were correctly retrieved, and after that the performance degraded quickly to 94% with 400,000 patterns, 71% with 460,000, and 23% with 550,000 (figure 4). Each run took about two hours of CPU time on a Cray Y-MP 8/864. From these simulations, and those with other configurations shown in figure 4, it can be concluded that the average capacity of the convergence-zone episodic model may be as much as one order of magnitude larger than the theoretical lower bound.

7 Less is More: The Effect of Sparse Connectivity

In the convergence-zone model so far, the feature maps have been fully connected to the binding layer. Such uniform configuration makes analysis and simulation of the model easier, but it is not very realistic. Both from the point of view of modeling the hippocampal memory and building artificial memory systems, it is important to know how the capacity is affected when the two layers are only sparsely connected.

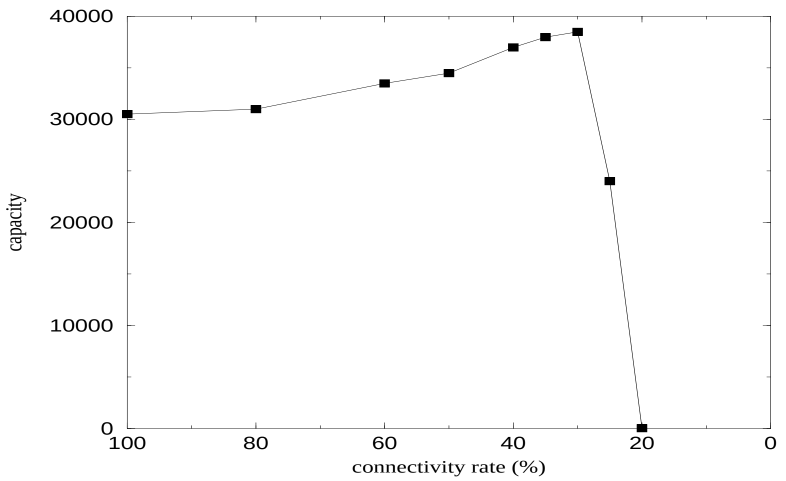


Figure 5: **Capacity with sparse connectivity.** The horizontal axis shows the percentage of connections that were available to form binding patterns, and the vertical axis indicates the capacity at the 99% confidence level. As connectivity decreases, the retrieval patterns become more focused, and the retrieval becomes more reliable until about 30% connectivity, where there are no longer enough connections to form binding patterns of m units. The model had $f = 1000, n = 3000, m = 20, t = 4, c = 3$, and the plot is an average of 5 simulations.

A series of simulations were run to determine how the model would perform with decreasing connectivity. A given percentage of connections were chosen randomly and disabled, and the 99% capacity point of the resulting model was found. The binding patterns were chosen slightly differently to account for missing connections: the m binding units were chosen among those binding layer units that were connected to all features in the feature pattern. If there were less than m such units, the binding pattern consisted all available units.

Due to high computational cost of the simulations, a small convergence-zone memory with $f = 1000$, $n = 3000$, $m = 20$, $t = 4$, and $c = 3$ was used in these experiments. At each level of connectivity from 100% down to 20%, five different simulations with different random connectivity were run and the results were averaged (figure 5). Since the main effect of sparse connectivity is to limit the number of binding units that are available for the binding pattern, one would expect that the capacity would go down. However, just the opposite turns out to be true: the fewer connections the model had available (downto 30% connectivity), the more patterns could be stored with 99% probability of correct retrieval.

The intuition turns out to be incorrect for an interesting reason: sparse connectivity causes the binding constellations to become more focused, removing spurious overlap that is the main cause of retrieval errors. To see this, consider how $\tilde{Z}$ (the size of the binding constellation of a feature unit after at least one pattern has been stored on it) grows when a new pattern is stored on the unit. A set of binding units is chosen to represent the pattern. When only a fraction r of the connections to the binding layer are available, there are only $r^t n$ binding units to choose from, compared to n in the fully connected model. It is therefore more likely that a chosen binding unit is already part of the binding constellation of the feature unit, and this constellation therefore grows at a slower pace than in the fully connected case. The expected size of the constellation after p patterns have been stored in the memory becomes (from equation 6)

$$\mathrm{E}(\tilde{Z}) = m + (r^t n - m)(1 - (1 - \frac{m}{r^t n f})^p), \tag{41}$$

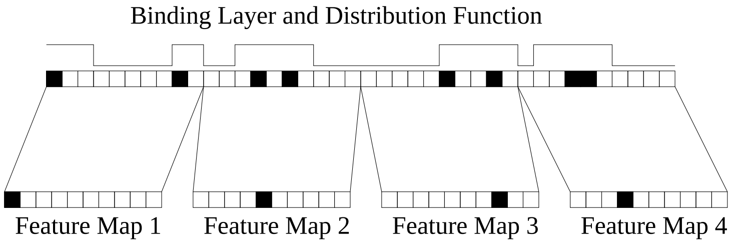


Figure 6: **Selecting a binding representation for an input pattern.** The binding layer is divided into sections. On the average m binding units will be selected from a distribution function that consists of one rectangular component for each section, centered around the unit whose location corresponds to the location of the input feature unit. If the center is close to the boundary of the section, the distribution component wraps around the section (as for Feature Map 1). The parameter σ determines the radius of the components, and thereby the variance in the binding pattern. This particular function had $\sigma = 2$, that is, the width of the individual components was $2 \times 2 + 1 = 5$. The model parameters were $f = 10, n = 40, m = 8$.

which decreases with connectivity r as long as there are at least m binding units available (i.e. $r^t n > m$). When the binding constellations are small, their intersections beyond the m common units will be small as well. During retrieval it is then less likely that a rogue unit will become more active than the correct unit. The activity patterns in the sparsely connected system are thus better focused, and retrieval more reliable.

When there are fewer than m binding units available, the capacity decreases very fast. In the model of figure 5 (with $m = 20$), the average number of binding units available is 45 for 35% connectivity, 24 for 30%, 12 for 25%, and 5 for 20%. In other words, the convergenze-zone episodic memory performs best when it is connected just enough to support activation of the sparse binding patterns. This is an interesting and surprising result, indicating that sparse connectivity is not just a necessity due to limited resources, but also gives a computational advantage. In the context of the hippocampal memory system it makes sense since evolution would be likely to produce a memory architecture that makes the best use of the available resources.

8 Error Behavior

When the binding patterns are selected at random as in the model so far, when errors occur during retrieval, the resulting feature values are also random. Human memory, however, rarely produces such random output. Human memory performance is often approximate, but robust and plausible. That is, when a feature cannot be retrieved exactly, a value is generated that is close to the correct value.

To test whether the convergence-zone architecture can model such "human-like" memory behavior, the storage mechanism must be changed so that it takes into account similarities between stored patterns. So far the spatial arrangement of the units has been irrelevant in the model; let us now treat the feature maps and the binding layer as one-dimensional maps (figure 6). The binding layer is divided into sections, one for each feature map. The binding units are selected stochastically from a distribution function that consists of one component for each of the sections. Each component is a flat rectangle with a radius of σ units around the unit whose location corresponds to the location of the active feature map unit. The distribution function is scaled so that on the average m units will be chosen for the entire binding pattern.

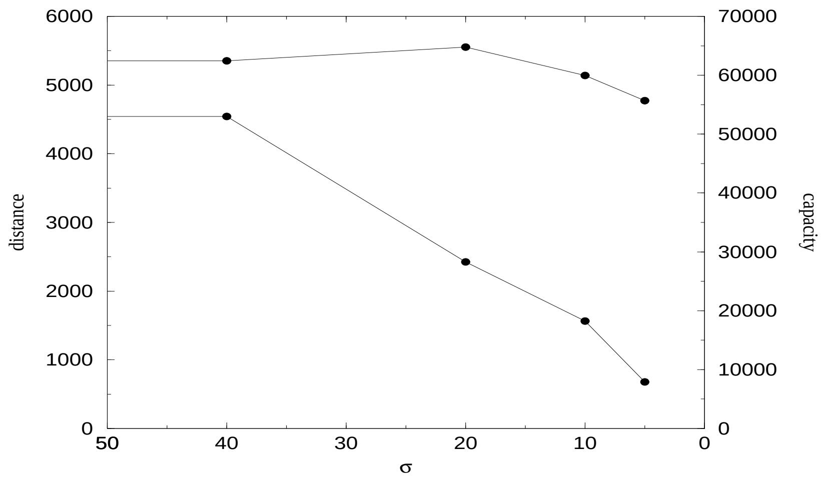


Figure 7: **Error behavior with descriptive binding patterns.** The lower curve (with scale at left) shows the average distance of incorrectly-retrieved features as a function of the distribution radius σ. The upper curve (with scale at right) indicates the corresponding capacity at the 99% confidence level. When the binding pattern is less random, the erroneous units tend to be closer to the correct ones. The model had $n = 20,000, f = 400, m = 20, t = 4, c = 3$. Retrieval of all stored patterns were checked. The plots indicate averages over 6 simulations.

The radius σ of the distribution components determines the variance of the resulting binding patterns. By setting $\sigma \geq (n/t - 1)/2$ a uniform distribution is obtained, which gives us the basic convergence-zone model. By making σ smaller, the binding representations for similar input patterns become more similar.

Several simulations with varying degrees of σ were run to see how the error behavior and the capacity of the model would be affected (figure 7). A configuration of 4 feature maps (3 cues) with $20,000$ binding units, 400 feature map units, and a 20-unit binding pattern was used. The results show that when the binding patterns are made more descriptive, the errors become more plausible: when an incorrect feature is retrieved, it tends to be close to the correct one. When the binding units are selected completely at random ($\sigma \geq (n/t - 1)/2$), the average distance is 5000; when $\sigma = 20$, it drops to 2500; and when $\sigma = 5$, to 700. Such "human-like" behavior is obtained with a slight cost in memory capacity. When the patterns become less random, there is more overlap in the encodings, and the capacity tends to decrease. This effect, however, appears to be rather minor, at least in the small models simulated in our experiments.

9 Discussion

The convergence-zone episodic model as analyzed and simulated above assumes that the feature patterns do not overlap much, and that the pattern is retrieved in a single iteration. Possibly relaxing these assumptions and the effects of such modifications are discussed below.

9.1 Pattern Overlap

The theoretical lower-bound calculations assumed that the chance of overlap of more than one feature is very small, and this was also true in the models that were analyzed and simulated.

However, this restriction does not limit the applicability of the model as much as it might first seem, for two reasons:

First, it might appear that certain feature values occur more often than others in the real world, causing more overlap than there currently is in the model. However, note that the input to the model is represented on feature maps. One of the basic properties of both computational and biological maps is that they adapt to the input distribution by magnifying the dense areas of the input space. In other words, if some perceptual experience is more frequent, more units will be allocated for representing it so that each unit gets to respond equally often to inputs (Kohonen 1989; Merzenich et al. 1984; Ritter 1991). Therefore, overlap in the feature map representations is significantly more rare than it may be in the absolute experience: the minor differences are magnified and the representations become more distinguishable and more memorable.

Second, as discussed in section 4.4, the chance of overlap of more than one feature is clearly small if the feature values are independent. For example in the coarse-grained model, at the 99% capacity point, on average there were 88 other patterns that shared exactly one common feature with a given pattern, whereas there were only 0.0078 other patterns that shared more than one feature. To be sure, in the real world the feature values across maps are correlated, which would make overlap of more than one feature more likely than it currently is in the model. While it is hard to estimate how common such correlations would be, they could grow quite a bit before they would become significant. In other words, the conclusions drawn from the current model are valid for at least small amounts of such correlations.

9.2 Progressive Recall

The retrieval process adopted in the convergence-zone model is a version of *simple recall* (Gardner-Medwin 1976), where the pattern is retrieved based on only direct associations from the retrieval cues. In contrast, *progressive recall* is an iterative process that uses the retrieved pattern at each step as the new retrieval cue. Progressive recall could be implemented in the convergence-zone model. Suppose features need to be retrieved in several maps. After the first retrieval attempt, the right feature unit will be clearly identified in most maps. For the second retrieval iteration, all these units can be used as cues, and it is likely that a pattern will be retrieved that is closer to the correct pattern than the one obtained with just simple recall. This way, progressive recall would cause an increase in the capacity of the model. Also, such a retrieval would probably be more robust against invalid retrieval cues (i.e. cues that are not part of the pattern to be retrieved). The dynamics of the progressive recall process are difficult to analyze (see Gibson and Robinson 1992 for a possible approach) and expensive to simulate, and simple recall was thus used in this first implementation of the convergence-zone model.

Above, a theoretical lower bound for the capacity of simple recall within a given error tolerance was derived, and the average capacity was estimated experimentally. Two other types of capacity can also be defined for an associative memory model (Amari 1988). The *absolute capacity* refers to the maximum number of patterns that the network can represent as equilibrium states, and the *relative capacity* is the maximum number of patterns that can be retrieved by progressive recall. The lower bound for the simple recall derived in this paper is also a lower bound for the absolute capacity, and thus also a lower bound for the relative capacity, which may be rather difficult to derive directly.

10 Related Work

Associative memory is one of the earliest and still most active areas of neural network research, and the convergence-zone model needs to be evaluated from this perspective. Although the architecture is mostly motivated by the neuroscience theory of perceptual maps, hippocampal encoding, and convergence zones, it is mathematically most closely related to statistical associative memories and the sparse distributed memory model. Contrasting the architecture with the Hopfield network and modified backpropagation is appropriate because these are the best-known associative memory models to date. Eventually convergence-zone memory might serve as a model of human episodic memory together with the trace feature map model described below. Although it is an abstract model of the hippocampal system, it is consistent with the more low-level models of the hippocampal circuitry, and complements them well.

10.1 The Hopfield model

The Hopfield network (Hopfield 1982) was originally developed to model the computational properties of neurobiological systems from the perspective of statistical mechanics (Amit et al. 1985a, 1985b; Kirkpatrick and Sherrington 1988; Peretto and Niez 1986). The Hopfield network is characterized by full connectivity, except from a unit to itself. Patterns can be stored one at a time, but the storage mechanism is rather involved. To store an additional pattern in a network of, say, N units, the weights of all the $N \times (N-1)$ connections have to be changed. In contrast, the convergence-zone memory is more sparse in that only $t \times m \ll n \ll f$ weights have to be modified.

A pattern is retrieved from the Hopfield network through progressive recall. The cues provide initial activation to the network, and the unit activations are updated asynchronously until they stabilize. The final stable activation pattern is then taken as the output of the network. In the convergence-zone model, on the other hand, retrieval is a four-step version of simple recall: first the activation is propagated from the input maps to the binding layer, thresholded, and then propagated back to all feature maps, where it is thresholded again. This algorithm can be seen as a computational abstraction of an underlying asynchronous process. In a more low-level implementation, thresholding would be achieved through inhibitory lateral connections. The neurons would update their activation one at a time in random order, and eventually stabilize to a state that represents retrieval of a pattern.

The capacity for the Hopfield network has been shown theoretically to be $N/4 \ln N$ (Amit 1989; Hertz et al. 1991; Keeler 1988; McEliece et al. 1986), and experimentally about $0.15N$ (Hopfield 1982). For the convergence-zone model such a simple closed-form formula is difficult to derive, because the model has many more parameters and correlations that complicate the analysis. However, as was shown above, a lower bound can be derived for a given set of parameters. Such bounds and also experimental simulations show that the capacity for the model can be orders of magnitude higher than the number of units, which is rather unique for an associative memory neural network.

However, it should be noted that in the convergence-zone model, each pattern is much smaller than the network. In a Hopfield network of size N each pattern contains N bits of information, while in the convergence-zone model each pattern consists of only t features. Each feature can be seen as a number between 1 and f, corresponding to its location in the feature map. To represent such a number, ${}^2\log f$ bits are needed, and a feature pattern thus contains $t\,{}^2\log f$ bits of information. Compared to the Hopfield model and other similar associative memory models, the information content of each pattern has been traded off for the capacity to store more of them in a network of equal size.

10.2 Backpropagation and Related Models

Several models of associative memory have been proposed that are based on backpropagation or similar incremental learning rules (Ackley et al. 1985; Anderson et al. 1977; Knapp and Anderson 1984; McClelland and Rumelhart 1986a, 1986b). However, these models suffer from catastrophic interference, which makes it difficult to apply them to modeling human associative memory (Grossberg 1987; McCloskey and Cohen 1989; Ratcliff 1990). If the patterns are to be learned incrementally, without repeating the earlier patterns, the later patterns in the sequence erase the earlier associations from memory.

Several techniques have been proposed to alleviate forgetting, including using weights with different learning rates (Hinton and Plaut 1987), gradually including new examples and phasing out earlier ones (Hetherington and Seidenberg 1989), forcing semidistributed hidden-layer representations (French 1991), concentrating changes on novel parts of the inputs (Kortge 1990), using units with localized receptive fields (Kruschke 1992), and adding new units and weights to encode new information (Fahlman 1991; Fahlman and Lebiere 1990). In these models, one-shot storage is still not possible, although the number of required iterations is reduced, and old information can be relearned very fast. At this point it is also unclear whether these architectures would scale up to number of patterns appropriate for human memory.

10.3 Statistical Associative Memory Models

The convergence-zone model is perhaps most closely related to the correlation matrix memory (Kohonen 1971, 1972; see also Anderson 1972; Cooper 1973). In this model there are a number of receptors (corresponding to feature maps in the convergence-zone model) that are connected to a set of associators (the binding layer). The receptors are divided into key fields where the retrieval cue is specified, and data fields where the retrieved pattern appears. Each key and data field corresponds to a feature map in the convergence-zone model. Instead of one value for each field, a whole feature map represents the field, modeling value-unit encoding in biological perceptual systems. There is no distinction between key and data fields either; every feature map can function as a key in the convergence-zone model.

Other related statistical models include the learning matrix (Steinbuch 1961) and the associative net (Willshaw et al. 1969), which are precursors of the correlation matrix model. These had a uniform matrix structure connecting inputs to outputs in a single step. Such models are simple and easy to implement in hardware, although they do not have a very high capacity (Faris and Maier 1988; Palm 1980, 1981). Other associative matrix models have relied on progressive recall, and therefore are similar in spirit to the Hopfield network. Little and Shaw's (1975) network of stochastic neurons, and Gardner-Medwin's (1976) and Marr's (1971) models of the hippocampus fall in this category. Progressive recall gives them a potentially higher capacity, which with high connectivity exceeds that of the Hopfield network (Gibson and Robinson 1992). They are also more plausible in that they do not require N^2 internal connections (where N is the number of units in the network), although the capacity decreases rapidly with decreasing connectivity.

10.4 Sparse Distributed Memory

The Sparse Distributed Memory model (SDM; Kanerva 1988) was originally developed as a mathematical abstraction of an associative memory machine. Keeler (1988) developed a neural-network implementation of the idea and showed that the SDM compares favorably with the Hopfield model; the capacity is larger and the patterns do not have to include all units in the network.

It is possible to give the convergence-zone memory model an interpretation as a special case of the SDM model: A fully-connected two-layer network consisting of a combined input/output layer and a hidden layer. In alternating steps, activity is propagated from the input/output layer to the hidden layer and back. Seen this way, every unit in the input/output layer corresponds to one feature map in the convergence-zone model, and the hidden layer corresponds to the binding layer.

It can be shown that the capacity of the SDM is independent of the size of the input/output layer. Moreover, if the size of the input/output layer is fixed, the capacity increases linearly with the size of the hidden layer. These results suggest that similar properties apply also to the convergence-zone episodic memory model.

10.5 Trace Feature Maps

The Trace Feature Map model of Miikkulainen (1992, 1993) consists of a self-organizing feature map where the space of all possible experiences is first laid out. The map is laterally fully connected with weights that are initially inhibitory. Traces of experiences are encoded as attractors using these lateral connections. When a partial pattern is presented to the map as a cue, the lateral connections move the activation pattern around the nearest attractor.

The Trace Map was designed as an episodic memory component of a story understanding system. The main emphasis was not on capacity, but on psychologically valid behavior. The basins of attraction for the traces interact, generating many interesting phenomena. For example, the later traces have larger attractor basins and are easier to retrieve, and unique traces are preserved even in an otherwise overloaded memory. On the other hand, because each basin is encoded through the lateral connections of several units, the capacity of the model is several times smaller than the number of units. Also, there is no mechanism for encoding truly novel experiences; only vectors that are already represented in the map can be stored. In this sense, the Trace Map model can be seen as the cortical component of the human long-term memory system. It is responsible for many of the effects, but incorporating novel experiences into its existing structure is a lengthy process, as it appears to be in human memory system (Halgren 1984; McClelland et al. 1995).

10.6 Models of the hippocampus

A large number of models of the hippocampal formation and its role in memory processing have been proposed (Alvarez and Squire 1994; Gluck and Myers 1993; McNaughton and Morris 1987; Marr 1971; Murre 1995; O'Reilly and McClelland 1994; Read et al. 1994; Schmajuk and DiCarlo 1992; Sutherland and Rudy 1989; Teyler and Discenna 1986; Treves and Rolls 1994, 1991; Wickelgren 1979). They include a more detailed description of the circuitry inside hippocampus, and aim at showing how memory traces could be created in such a circuitry. The convergence-zone model operates at a higher level of abstraction than these models, and in this sense is complementary to them.

Marr (1971) presented a detailed theory of the hippocampal formation, including numerical constraints, capacity analysis, and interpretation at the level of neural circuitry. The input was based on local input fibers from the neocortex, processed by an input and output layer of hippocampal neurons with collateral connections. Recall was based on recurrent completion of a pattern. The convergence-zone memory can be seen as a high-level abstration of Marr's theory, with the emphasis on the convergence-zone structure that allows for higher capacity than Marr predicted.

Several authors have proposed a role for the hippocampus similar to the convergence-zone idea (Alvarez and Squire 1994; McClelland et al. 1995; Murre 1995; Teyler and Discenna 1986; Treves and Rolls 1994; Wickelgren 1979). In these models, hippocampus itself does not store a complete representation of the episode, but acts as an indexing area, or compressed representation, that binds together parts of the actual representation in the neocortical areas. Treves and Rolls (1994) also propose backprojection circuitry for accomplishing such recall. Convergenze-zone memory is consistent with such interpretations, focusing on analyzing the capacity of such structures.

One of the assumptions of the convergence-zone memory, motivated by recent results by Wilson and McNaughton (1993), is that the binding encoding is sparse and random. The model by O'Reilly and McClelland (1994) shows how the hippocampal circuitry could form such sparse, diverse encodings. They explored the tradeoff between pattern separation and completion and showed that the circuitry could be set up to perform both of these tasks simultaneously. The entorhinal–dentate–CA3 pathway could be responsible for forming random endocings of traces in CA3, and the separation between storage and recall could be due to overall difference in the activity level in the system.

11 Future Work

The future work on the convergenze-zone episodic memory model will focus on three areas. First, the model can be extended in several ways towards a more accurate model of the actual neural processes. For instance, lateral inhibitory connections between units within a feature map could be added to select the unit with the highest activity. A similar extension could be applied to the binding layer; only instead of a single unit, multiple units should stay active in the end. Lateral connections in the binding layer could also be used to partially complete the binding pattern even before propagation to the feature maps. As the next step, the binding layer could be expanded to take into account finer structure in the hippocampus, including the encoding and retrieval circuitry proposed by O'Reilly and McClelland (1994). A variation of the Hebbian learning mechanism (Hebb 1949; Miller and MacKay 1992) could then be used to implement the storage and recall mechanisms. In addition to providing insight to the hippocampal memory system, such research could lead to a practical implementation of the convergence zone memory, and perhaps even to a hardware implementation.

Second, a number of potential extensions to the model could be studied in more detail. It might be possible to take sparse connectivity into account in the analysis, and obtain tighter lower bounds in this more realistic case. Recurrency could be introduced between feature maps and the binding layer, and capacity could be measured under progressive recall. Possibilities for extending the model to tolerate more overlap between patterns should also be studied.

Third, the model could be used as a stepping stone towards a more comprehensive model of human episodic memory, including modules for the hippocampus and the neocortical component. As discussed above, the convergence-zone model seems to fit the capabilities of the hippocampal component well, whereas something like trace feature maps could be used to model the cortical component. It would be necessary to observe and characterize the memory interference effects of the components and compare them with experimental results on human memory. However, the main challenge of this research would be on the interaction of the components, that is, how the hippocampal memory could transfer its contents to the cortical component. At this point it is not clear how this process could take place, although several proposals exist (Alvarez and Squire 1994; Halgren 1984; McClelland et al. 1995; Milner 1989; Murre 1995; Treves and Rolls 1994). Computational investigations could prove instrumental in understanding the foundations of this remarkable system.

12 Conclusion

Mathematical analysis and experimental simulations show that a large number of episodes can be stored in the convergence-zone memory with reliable content-addressable retrieval. For the hippocampus, a sufficient capacity can be achieved with a fairly small number of units and connections. Moreover, the convergence zone itself requires only a fraction of the hardware required for perceptual representation. These results provide a possible explanation for why human memory is so efficient with such a high capacity, and why memory areas appear small compared to the areas devoted to low-level perceptual processing. It also suggests that the computational units of the hippocampus and the perceptual maps can be quite coarse, and gives a computational reason why the maps and the hippocampus should be sparsely connected.

The model makes use of the combinatorics and the clean-up properties of coarse coding in a neurally-inspired architecture. The practical storage capacity of the model appears to be at least two orders of magnitude higher than that of the Hopfield model with the same number of units, while using two orders of magnitude fewer connections. On the other hand, the patterns in the convergence-zone model are smaller than in the Hopfield network. Simulations also show that psychologically valid error behavior can be achieved if the binding patterns are made more descriptive: the erroneous patterns are close to the correct ones. The convergence-zone episodic memory is a step towards a psychologically and neurophysiologically accurate model of human episodic memory, the foundations of which are only now beginning to be understood.

References

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, 9:147–169.

Alon, N., and Spencer, J. H. (1992). *The Probabilistic Method.* New York: Wiley.

Alvarez, P., and Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: A simple network model. *Proceedings of the National Academy of Sciences of the USA*, 91:7041–7045.

Amaral, D. G., Ishizuka, N., and Claiborne, B. (1990). Neurons, numbers and the hippocampal network. In Storm-Mathisen, J., Zimmer, J., and Ottersen, O. P., editors, *Understanding the Brain through the Hippocampus*, vol. 83 of *Progress in Brain Research*, 1–11. New York: Elsevier.

Amari, S.-I. (1977). Neural theory of association and concept formation. *Biological Cybernetics*, 26:175–185.

Amari, S.-I. (1988). Associative memory and its statistical neurodynamical analysis. In Haken, H., editor, *Neural and Synergetic Computers*, 85–99. Berlin: Springer Verlag.

Amit, D. J. (1989). *Modeling Brain Function: The World of Attractor Neural Networks.* Cambridge, UK: Cambridge University Press.

Amit, D. J., Gutfreund, H., and Sompolinsky, H. (1985a). Spin-glass models of neural networks. *Physical Review A*, 32:1007–1018.

Amit, D. J., Gutfreund, H., and Sompolinsky, H. (1985b). Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, 55:1530–1533.

Anderson, J. A. (1972). A simple neural network generating an interactive memory. *Mathematical Biosciences*, 14:197–220.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., and Jones, R. S. (1977). Distinctive features, categorical perception and probability learning: Some applications of a neural model. *Psychological Review*, 84:413–451.

Bain, L. J., and Engelhardt, M. (1987). *Introduction to Probability and Mathematical Statistics.* Boston, MA: PWS Publishers.

Boss, B., Peterson, G., and Cowan, W. (1985). On the numbers of neurons in the dentate gurys of the rat. *Brain Research*, 338:144–150.

Boss, B., Turlejski, K., Stanfield, B., and Cowan, W. (1987). On the numbers of neurons in fields CA1 and CA3 of the hippocampus of Sprague-Dawley and Wistar rats. *Brain Research*, 406:280–287.

Cooper, L. N. (1973). A possible organization of animal memory and learning. In Lundquist, B., and Lundquist, S., editors, *Proceedings of the Nobel Symposium on Collective Properties of Physical Systems*, 252–264. New York: Academic Press.

Damasio, A. R. (1989a). The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation*, 1:123–132.

Damasio, A. R. (1989b). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, 33:25–62.

Fahlman, S. E. (1991). The recurrent cascade-correlation architecture. In Lippmann, R. P., Moody, J. E., and Touretzky, D. S., editors, *Advances in Neural Information Processing Systems 3*, 190–205. San Mateo, CA: Morgan Kaufmann.

Fahlman, S. E., and Lebiere, C. (1990). The cascade-correlation learning architecture. In Touretzky, D. S., editor, *Advances in Neural Information Processing Systems 2*, 524–532. San Mateo, CA: Morgan Kaufmann.

Faris, W. G., and Maier, R. S. (1988). Probabilistic analysis of a learning matrix. *Advances in Applied Probability*, 20:695–705.

French, R. M. (1991). Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks. In *Proceedings of the 13th Annual Conference of the Cognitive Science Society*, 173–178. Hillsdale, NJ: Erlbaum.

Gardner-Medwin, A. R. (1976). The recall of events through the learning of associations between their parts. *Proceedings of the Royal Society of London B*, 194:375–402.

Gibson, W. G., and Robinson, J. (1992). Statistical analysis of the dynamics of a sparse associative memory. *Neural Networks*, 5:645–661.

Gluck, M. A., and Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*, 3:491–516.

Grossberg, S. (1983). *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition and Motor Control.* Dordrecht, Holland; Boston: Reidel.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, 11:23–63.

Halgren, E. (1984). Human hippocampal and amygdala recording and stimulation: Evidence for a neural model of recent memory. In Squire, L., and Butters, N., editors, *The Neuropsychology of Memory*, 165–182. New York: Guilford.

Hasselmo, M. E., Rolls, E. T., and Baylis, G. C. (1989). The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behavioural Brain Research*, 32:203–218.

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory.* New York: Wiley.

Heit, G., Smith, M. E., and Halgren, E. (1989). Neural encoding of individual words and faces by the human hippocampus and amygdala. *Nature*, 333:773–775.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation.* Reading, MA: Addison-Wesley.

Hetherington, P. A., and Seidenberg, M. S. (1989). Is there "catastrophic interference" in connectionist networks? In *Proceedings of the 11th Annual Conference of the Cognitive Science Society*, 26–33. Hillsdale, NJ: Erlbaum.

Hinton, G. E., and Anderson, J. A., editors (1981). *Parallel Models of Associative Memory.* Hillsdale, NJ: Erlbaum.

Hinton, G. E., and Plaut, D. C. (1987). Using fast weights to deblur old memories. In *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, 177–186. Hillsdale, NJ: Erlbaum.

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

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences, USA*, 81:3088–3092.

Jessell, E. R. K. (1991). Nerve cells and behavior. In Kandel, E. R., Schwartz, J. H., and Jessell, T. M., editors, *Principles of Neural Science*, 18–32. Elsevier.

Kairiss, E. W., and Miranker, W. L. (1996). Cortical memory dynamics. Technical Report 9604, Neuroengineering and Neuroscience Center, Yale University, New Haven, CT.

Kanerva, P. (1988). *Sparse Distributed Memory.* Cambridge, MA: MIT Press.

Keeler, J. D. (1988). Comparison between Kanerva's SDM and Hopfield-type neural networks. *Cognitive Science*, 12:299–329.

Kirkpatrick, S., and Sherrington, D. (1988). Infinite range models of spin-glasses. *Physical Review B*, 17:4384–4403.

Knapp, A. G., and Anderson, J. A. (1984). Theory of categorization based on distributed memory storage. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 10:616–637.

Knudsen, E. I., du Lac, S., and Esterly, S. D. (1987). Computational maps in the brain. In Cowan, W. M., Shooter, E. M., Stevens, C. F., and Thompson, R. F., editors, *Annual Review of Neuroscience*, 41–65. Palo Alto: Annual Reviews.

Kohonen, T. (1971). A class of randomly organized associative memories. *Acta Polytechnica Scandinavica*, EL 25.

Kohonen, T. (1972). Correlation matrix memories. *IEEE Transactions on Computers*, C-21:353–359.

Kohonen, T. (1977). *Associative Memory: A System-Theoretical Approach.* Berlin; Heidelberg; New York: Springer.

Kohonen, T. (1989). *Self-Organization and Associative Memory.* Berlin; Heidelberg; New York: Springer. Third edition.

Kohonen, T., and Mäkisara, K. (1986). Representation of sensory information in self-organizing feature maps. In Denker, J. S., editor, *Neural Networks for Computing*, 271–276. New York: American Institute of Physics.

Kortge, C. A. (1990). Episodic memory in connectionist networks. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, 764–771. Hillsdale, NJ: Erlbaum.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99:22–44.

Little, W. A., and Shaw, G. L. (1975). A statistical theory of short and long term memory. *Behavioral Biology*, 14:115–133.

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London B*, 262:23–81.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102:419–457.

McClelland, J. L., and Rumelhart, D. E. (1986a). Amnesia and distributed memory. In Rumelhart, D. E., and McClelland, J. L., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 2: Psychological and Biological Models*, 503–527. Cambridge, MA: MIT Press.

McClelland, J. L., and Rumelhart, D. E. (1986b). A distributed model of human learning and memory. In McClelland, J. L., and Rumelhart, D. E., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 2: Psychological and Biological Models*, 170–215. Cambridge, MA: MIT Press.

McCloskey, M., and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation*, 24:104–169.

McEliece, R. J., Posner, E. C., Rodemich, E. R., and Venkatesh, S. S. (1986). The capacity of the hopfield associative memory. *IEEE Transactions on Information Theory*, 33:461–482.

McNaughton, B. L., and Morris, R. G. M. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neuroscience*, 10:408–415.

Merzenich, M. M., Nelson, R. J., Stryker, M. P., Cynader, M. S., Schoppmann, A., and Zook, J. M. (1984). Somatosensory cortical map changes following digit amputation in adult monkeys. *Journal of Comparative Neurology*, 224:591–605.

Miikkulainen, R. (1992). Trace feature map: A model of episodic associative memory. *Biological Cybernetics*, 67:273–282.

Miikkulainen, R. (1993). *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory.* Cambridge, MA: MIT Press.

Miller, K. D., and MacKay, D. J. C. (1992). The role of constrains in Hebbian learning. CNS Memo 19, Computation and Neural Systems Program, California Institute of Technology, Pasadena, CA.

Milner, P. (1989). A cell assembly theory of hippocampal amnesia. *Neuropshychologia*, 27:23–30.

Murre, J. M. J. (1995). A model of amnesia. Manuscript.

O'Reilly, R. C., and McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus*, 4:661–682.

Palm, G. (1980). On associative memory. *Biological Cybernetics*, 36:19–31.

Palm, G. (1981). On the storage capacity of an associative memory with randomly distributed storage elements. *Biological Cybernetics*, 39:125–127.

Peretto, P., and Niez, J. (1986). Collective properties of neural networks. In Bienenstock, E., Fogelman Soulié, F., and Weisbuch, G., editors, *Disordered Systems and Biological Organization*, 171–185. Berlin; Heidelberg; New York: Springer.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, 97:285–308.

Read, W., Nenov, V. I., and Halgren, E. (1994). Role of inhibition in memory retrieval by hippocampal area CA3. *Neuroscience and Biobehavioral Reviews*, 18:55–68.

Ritter, H. J. (1991). Asymptotic level density for a class of vector quantization processes. *IEEE Transactions on Neural Networks*, 2:173–175.

Rolls, E. T. (1984). Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. *Human Neurobiology*, 3:209–222.

Schmajuk, N. A., and DiCarlo, J. J. (1992). Stimulus configuration, classical conditioning, and hippocampal function. *Psychological Review*, 99:268–305.

Sejnowski, T. J., and Churchland, P. S. (1989). Brain and cognition. In Posner, M. I., editor, *Foundations of Cognitive Science*, chapter 8, 315–356. Cambridge, MA: MIT Press.

Squire, L. R. (1987). *Memory and Brain*. Oxford, UK; New York: Oxford University Press.

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

Squire, L. R., Shimamura, A. P., and Amaral, D. G. (1989). Memory and the hippocampus. In Byrne, J. H., and Berry, W. O., editors, *Neural Models of Plasticity*, 208–239. New York: Academic Press.

Steinbuch, K. (1961). Die lernmatrix. *Kybernetic*, 1:36–45.

Sutherland, R. W., and Rudy, J. W. (1989). Configural association theory: The role of the hippocampal formation in learning, memory and amnesia. *Psychobiology*, 17:129–144.

Teyler, T. J., and Discenna, P. (1986). The hippocampal memory indexing theory. *Behavioral Neuroscience*, 100:147–154.

Treves, A., and Rolls, E. T. (1991). What determines the capacity of autoassociative memories in the brain? *Network*, 2:371–398.

Treves, A., and Rolls, E. T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus*, 4:374–391.

Tulving, E. (1972). Episodic and semantic memory. In Tulving, E., and Donaldson, W., editors, *Organization of Memory*, 381–403. New York: Academic Press.

Tulving, E. (1983). *Elements of Episodic Memory*. Oxford, UK; New York: Oxford University Press.

Wickelgren, W. A. (1979). Chunking and consolidation: A theoretical synthesis of semantic networks, configuring, S-R versus cognitive learning, normal forgetting, the amnesic syndrome, and the hippocampal arousal system. *Psychological Review*, 86:44–60.

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

Wilson, M. A., and McNaughton, B. L. (1993). Dynamics of the hippocampal ensemble code for space. *Science*, 261:1055–1058.

A Probability Theory Background and Proofs

In this appendix, concepts from probability theory that are necessary for understanding the main text are briefly reviewed, and details of the probabilistic formulation and martingale analysis are presented (for more background on probability theory and statistics see e.g. Alon and Spencer 1992, or Bain and Engelhardt 1987).

A.1 Distributions and bounds

In the analysis of sections 3 and 4, two probability density functions are used. The first one is the binomial distribution with parameters n and p, denoted as $\mathrm{B}(n,p)$, and it gives the outcome of n independent trials where the two possible alternatives have probability p and $1-p$. This distribution has the expected value np and variance $np(1-p)$. The other one is the hypergeometric distribution with parameters n, m and N, denoted as $\mathrm{HYP}(n,m,N)$, and representing the number of elements in common between two independently-chosen subsets of n and m elements of a common superset of N elements. The distribution has the expected value $\frac{nm}{N}$ and variance $\frac{nm}{N}(1-\frac{m}{N})\frac{N-n}{N-1}$.

The Chernoff bounds can be used to estimate how likely a binomially distributed variable, X, is to have a value within a given distance δ from its mean:

$$\mathrm{P}(X \leq (1-\delta)np) \leq \left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{np}, \qquad 0<\delta<1, \tag{42}$$

$$\mathrm{P}(X \geq (1+\delta)np) \leq \left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{np}, \qquad \delta>0. \tag{43}$$

Even if the trials are not independent (and therefore X is not binomially distributed), in some cases the sequence of variables $X_0,\ldots,X_n$ can be analyzed as a martingale, and bounds similar to Chernoff bounds derived using Azuma's inequalities (see appendix A.3).

A.2 Details of the probabilistic formulation

As in section 3, let Z_i be the size of the binding constellation of a feature unit after i patterns have been stored on it, and let Y_i be its increase after storing the ith pattern on it. Then

$$Z_i = \sum_{k=1}^{i} Y_k. \tag{44}$$

Let $Y_i, i>1$ be hypergeometrically distributed with parameters m, $n-z_{i-1}$, and n:

$$\mathrm{P}(Y_i = y | Z_{i-1} = z_{i-1}) = \binom{n-z_{i-1}}{y}\binom{z_{i-1}}{m-y} \Big/ \binom{n}{m}, \tag{45}$$

and let $Z_1 = Y_1 = m$ with probability 1. Then

$$\begin{aligned}
\mathrm{E}(Y_i) &= \mathrm{E}(\frac{m(n-Z_{i-1})}{n}) \\
&= m - \frac{m}{n}\mathrm{E}(Z_{i-1})
\end{aligned}$$

$$
\begin{aligned}
&= m - \frac{m}{n}\sum_{k=1}^{i-1} \mathrm{E}(Y_k) \\
&= (1-\frac{m}{n})\mathrm{E}(Y_{i-1}) \\
&= m(1-\frac{m}{n})^{i-1}. \qquad (46)
\end{aligned}
$$

Using equation 46, an expression for $\mathrm{E}(Z_i)$ can be derived:

$$
\begin{aligned}
\mathrm{E}(Z_i) &= \sum_{k=1}^{i} \mathrm{E}(Y_k) \\
&= \sum_{k=1}^{i} m(1-\frac{m}{n})^{k-1} \\
&= n(1-(1-\frac{m}{n})^{i}). \qquad (47)
\end{aligned}
$$

This equation indicates that initially, when no patterns are stored, $Z_i = 0$, and that Z_i converges to n as i goes to infinity.

The variance of Z_i can be computed in a similar way. First a recurrent expression for the variance of Y_i is derived:

$$
\begin{aligned}
\mathrm{Var}(Y_i) &= \mathrm{E}_{Z_{i-1}}[\mathrm{Var}(Y_i|Z_{i-1})] + \mathrm{Var}_{Z_{i-1}}[\mathrm{E}(Y_i|Z_{i-1})] \\
&= \mathrm{E}_{Z_{i-1}}[m\tfrac{n-Z_{i-1}}{n}(1-\tfrac{n-Z_{i-1}}{n})\tfrac{n-m}{n-1}] + \mathrm{Var}_{Z_{i-1}}[m\frac{n-Z_{i-1}}{n}] \\
&= \tfrac{m(n-m)}{n-1}\left(1-(1-\tfrac{m}{n})^{i-1}\right)\left(1-\tfrac{m}{n}\right)^{i-1} + \tfrac{m(m-1)}{n(n-1)}\mathrm{Var}(Z_{i-1}). \qquad (48)
\end{aligned}
$$

To obtain the variance of Z_i, the covariance of Z_{i-1} and Y_i needs to be computed:

$$
\begin{aligned}
\mathrm{Cov}(Z_{i-1}, Y_i) &= \mathrm{E}(Z_{i-1}Y_i) - \mathrm{E}(Z_{i-1})\mathrm{E}(Y_i) \\
&= \mathrm{E}_{Z_{i-1}}[\mathrm{E}(Z_{i-1}Y_i|Z_{i-1})] - \mathrm{E}(Z_{i-1})\mathrm{E}(Y_i) \\
&= \mathrm{E}(Z_{i-1}\frac{m(n-Z_{i-1})}{n}) - \mathrm{E}(Z_{i-1})\mathrm{E}(Y_i) \\
&= m\mathrm{E}(Z_{i-1}) - \frac{m}{n}\left[(\mathrm{E}(Z_{i-1}))^2 + \mathrm{Var}(Z_{i-1})\right] - \mathrm{E}(Z_{i-1})\mathrm{E}(Y_i) \\
&= -\frac{m}{n}\mathrm{Var}(Z_{i-1}). \qquad (49)
\end{aligned}
$$

The variance of Z_i can now be easily derived:

$$
\begin{aligned}
\mathrm{Var}(Z_i) &= \mathrm{Var}(Z_{i-1} + Y_i) \\
&= \mathrm{Var}(Z_{i-1}) + \mathrm{Var}(Y_i) + 2\mathrm{Cov}(Z_{i-1}, Y_i) \\
&= \sum_{k=1}^{i} \mathrm{Var}(Y_k) + 2\sum_{k=2}^{i} \mathrm{Cov}(Z_{k-1}, Y_k) \\
&= \sum_{k=1}^{i} \mathrm{Var}(Y_k) - \frac{2m}{n}\sum_{k=1}^{i-1} \mathrm{Var}(Z_k) \\
&= \mathrm{Var}(Y_i) + (1-\frac{2m}{n})\mathrm{Var}(Z_{i-1}) \\
&= n(1-\frac{m}{n})^{i}(1-n(1-\frac{m}{n})^{i}) + n(n-1)\left(1-\tfrac{m(2n-m-1)}{n(n-1)}\right)^{i}. \qquad (50)
\end{aligned}
$$

The base of every exponential in this expression is smaller than 1, so the variance goes to 0 as i goes to infinity. This makes sense in the model: If many patterns are stored, it becomes very likely that Z_i has a value close to n, and hence its variance should go to zero.

So far it has been assumed that i is an ordinary variable. If it is replaced by a stochastic variable I that is binomially distributed with parameters p and $\frac{1}{f}$, the previously computed expected values and variances must be made conditional on I. To compute the unconditional expected values and variances, the following lemma is needed:

Lemma A.1 *If* $X \sim \mathrm{B}(n,p)$, *then* $\mathrm{E}\left((1-a)^X\right) = (1-ap)^n$.

Proof:

$$\begin{aligned}
\mathrm{E}\left((1-a)^X\right) &= \sum_{x=0}^{n}(1-a)^x \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=0}^{n} \binom{n}{x} (p-ap)^x (1-p)^{n-x} \\
&= ((p-ap)+(1-p))^n \\
&= (1-ap)^n \qquad \Box.
\end{aligned}$$

Let Z denote the unconditional value of Z_i. The desired results follow immediately from lemma A.1 and the linearity of expected values:

$$\begin{aligned}
\mathrm{E}(Z) &= n(1-(1-\frac{m}{nf})^p), &(51)\\
\mathrm{Var}(Z) &= \mathrm{Var}_I[\mathrm{E}(Z_I|I)] + \mathrm{E}_I[\mathrm{Var}(Z_I|I)] \\
&= n^2 \mathrm{Var}\left[(1-\tfrac{m}{n})^I\right] \\
&\quad + \mathrm{E}\left[n(1-\tfrac{m}{n})^I - n^2(1-\tfrac{m}{n})^{2I} + n(n-1)(1-\tfrac{m(2n-m-1)}{n(n-1)})^I\right] \\
&= n(1-\tfrac{m}{nf})^p(1-n(1-\tfrac{m}{nf})^p) + n(n-1)\left(1-\tfrac{m(2n-m-1)}{n(n-1)f}\right)^p. &(52)
\end{aligned}$$

Let $\tilde{Z}$ be defined as $\tilde{Z} \sim Z_I | I \geq 1$. Then $\tilde{Z}$ will always be at least m. Expressions similar to equations 51 and 52 can be derived for $\tilde{Z}$:

$$\begin{aligned}
\mathrm{E}(\tilde{Z}) &= m + (n-m)(1-(1-\frac{m}{nf})^p), &(53)\\
\mathrm{Var}(\tilde{Z}) &= (n-m)(1-\tfrac{m}{nf})^{p-1}(1-(n-m)(1-\tfrac{m}{nf})^{p-1}) \\
&\quad + (n-m)(n-m-1)\left(1-\tfrac{m(2n-m-1)}{n(n-1)f}\right)^{p-1}. &(54)
\end{aligned}$$

From equations 51 and 53 it follows that on the average, $\tilde{Z}$ is larger than Z. Initially the difference is exactly m, and the difference goes to zero as p goes to infinity. This means that initially, when few patterns are stored, the right units are very likely to be retrieved, since they have the advantage of receiving activation from at least m binding units. However, when more patterns are stored, almost every unit in a feature map receives the same amount of activation. Units that have been used most often are most likely to be retrieved.

It is difficult to derive an exact expression for the expected value and variance of X_c (the intersection of c retrieval cues) because the binding constellations are correlated. Since the same partial patterns might have been stored several times, certain combinations of retrieval cues can give rise to spurious activity in the binding layer, which is difficult to take into account in the analysis. For this reason, the analysis is carried out under the assumption that the chance of partial patterns is negligible, which is reasonable in most cases and easy to check.

A.3 Martingales

Martingales give us a way to analyze the outcome of n trials with limited dependencies. A martingale is a sequence of random variables $X_0, \ldots, X_n$ such that

$$\mathrm{E}(X_i|X_0, \ldots, X_{i-1}) = X_{i-1}, \quad 1 \le i \le n. \tag{55}$$

If a martingale satisfies the Lipschitz condition, that is

$$|X_i - X_{i-1}| \le 1, \quad 1 \le i \le n, \tag{56}$$

then the following inequalities hold:

$$\mathrm{P}(X_i \le X_0 - \lambda\sqrt{i}) \le e^{\frac{-\lambda^2}{2}} \tag{57}$$

$$\mathrm{P}(X_i \le X_0 + \lambda\sqrt{i}) \le e^{\frac{-\lambda^2}{2}}. \tag{58}$$

These equations are called Azuma's inequalities and they can be used to bound the final value of the sequence within a chosen confidence level.

A.4 The binding constellation martingale

Let Z_v^{E} be defined as in section 4.2:

$$\begin{aligned} Z_v^{\mathrm{E}} &= Z'_v + (n - Z'_v)(1 - (1 - \frac{1}{n})^{ki-v}) \\ &= n - (n - Z'_v)(1 - \frac{1}{n})^{ki-v}. \end{aligned} \tag{59}$$

In order to apply Azuma's inequality, the Lipschitz condition $|Z_v^{\mathrm{E}} - Z_{v-1}^{\mathrm{E}}| \le 1$ must to be satisfied. The binding units are chosen one at a time, and they may either already be part of the constellation or add one more unit to it. Therefore, the difference between Z'_v and Z'_{v-1} is always either 0 or 1:

Case 1: $Z'_v - Z'_{v-1} = 0$

$$\begin{aligned} |Z_v^{\mathrm{E}} - Z_{v-1}^{\mathrm{E}}| &= |-(n - Z'_v)(1 - \frac{1}{n})^{ki-v}(1 - (1 - \frac{1}{n}))| \\ &= |-\frac{n - Z'_v}{n}(1 - \frac{1}{n})^{ki-v}| \\ &\le 1. \end{aligned} \tag{60}$$

Case 2: $Z'_v = Z'_{v-1} + 1$

$$|Z_v^{\mathrm{E}} - Z_{v-1}^{\mathrm{E}}| = |-(n - Z'_v)(1 - \frac{1}{n})^{ki-v}(1 - (1 - \frac{1}{n})) + (1 - \frac{1}{n})^{ki-v+1}|$$

$$
\begin{aligned}
&= |-\frac{n-Z'_v}{n}(1-\frac{1}{n})^{ki-v} + (1-\frac{1}{n})^{ki-v+1}| \\
&= |(\frac{n-1}{n} - \frac{n-Z'_v}{n})(1-\frac{1}{n})^{ki-v}| \\
&= |\frac{Z'_v - 1}{n}(1-\frac{1}{n})^{ki-v}| \\
&\leq 1. && (61)
\end{aligned}
$$

In both cases we have $|Z_v^{\mathrm{E}} - Z_{v-1}^{\mathrm{E}}| \leq 1$ and hence Azuma's inequality can be applied to obtain a bound for Z.

A.5 The intersection martingale

Let X_v^{E} be defined as in section 4.3:

$$
\begin{aligned}
X_v^{\mathrm{E}} &= X'_v + \frac{(n_1 - v)(n_2 - X'_v)}{n_s - v} \\
&= \frac{n_s - n_1}{n_s - v} X'_v + \frac{(n_1 - v)n_2}{n_s - v}. && (62)
\end{aligned}
$$

First, $X_0^{\mathrm{E}}, \ldots, X_{n_1}^{\mathrm{E}}$ is shown to be a martingale. The expected increase of X'_v is $\frac{n_2 - X'_{v-1}}{n_s - v + 1}$. The conditional expected value of X_v^{E} given X_{v-1}^{E} is

$$
\begin{aligned}
\mathrm{E}(X_v^{\mathrm{E}} | X_{v-1}^{\mathrm{E}} = x_{v-1}^{\mathrm{E}}) &= \mathrm{E}(X_v^{\mathrm{E}} | X'_{v-1} = x'_{j-1}) \\
&= \frac{n_s - n_1}{n_s - v}(x'_{j-1} + \frac{n_2 - x'_{j-1}}{n_s - v + 1}) + \frac{(n_1 - v)n_2}{n_s - v} \\
&= \frac{n_s - n_1}{n_s - v + 1} x'_{j-1} + \frac{(n_s - n_1)n_2}{(n_s - v)(n_s - v + 1)} + \frac{(n_1 - v)n_2}{n_s - v} \\
&= \frac{n_s - n_1}{n_s - v + 1} x'_{j-1} + \frac{(n_s - n_1)n_2}{(n_s - v)(n_s - v + 1)} + \frac{(n_s - v)(n_1 - v)n_2 + (n_1 - v)n_2}{(n_s - v)(n_s - v + 1)} \\
&= \frac{n_s - n_1}{n_s - v + 1} x'_{j-1} + \frac{(n_1 - v + 1)n_2}{n_s - v + 1} \\
&= x_{v-1}^{\mathrm{E}}. && (63)
\end{aligned}
$$

Since $\mathrm{E}(X_v^{\mathrm{E}} | X_{v-1}^{\mathrm{E}}) = X_{v-1}^{\mathrm{E}}$, the sequence $X_0^{\mathrm{E}}, \ldots, X_{n_1}^{\mathrm{E}}$ is a martingale.

To apply Azuma's inequality, it is necessary to show that $|X_v^{\mathrm{E}} - X_{v-1}^{\mathrm{E}}| \leq 1$ for $v = 1, \ldots, n_1$. Unlike for the martingale of section A.4, this is not generally true, but true only when certain constraints on n_1, n_2 and n_s are satisfied. If these parameters are fixed, the difference can be seen as a function of v, X'_v and X'_{v-1}. The function is monotonic in these variables, and to find its maximum, it is sufficient to look only at the extreme values of v, X'_v and X'_{v-1}. The maximum in turn determines the constraints on n_1, n_2 and n_s.

As in section A.4, the proof consists of two main cases: (1) $X'_v = X'_{v-1}$ and (2) $X'_v = X'_{v-1} + 1$. Each of these is divided into several subcases. The following tables show the cases, the absolute difference in each case between X_v^{E} and X_{v-1}^{E}, and the constraint that follows from the difference.

Case 1: $X'_v = X'_{v-1}$.

In this case, using equation 62, $|X_v^{\mathrm{E}} - X_{v-1}^{\mathrm{E}}|$ can be rewritten as $\frac{(n_2 - X'_v)(n_s - n_1)}{(n_s - v)(n_s - v + 1)}$. There are three subcases:

Subcase	Difference	Resulting constraint
$v=1, X'_v=0$	$\frac{n_2(n_s-n_1)}{n_s(n_s-1)}$	$n_1>0 \vee n_2<n_s$
$v=n_1, X'_v=\max(0,n_1+n_2-n_s)$	$\frac{n_2}{n_s-n_1+1}$ if $X'_v=0$ $\frac{n_s-n_1}{n_s-n_1+1}$ if $X'_v=n_1+n_2-n_s$	$n_1+n_2-1\le n_s$ no constraint
$v=n_1, X'_v=\min(n_1,n_2)$	$\frac{n_2-n_1}{n_s-n_1+1}$ if $X'_v=n_1$ 0 if $X'_v=n_2$	no constraint

Case 2: $X'_v = X'_{v-1}+1$.

Now $|X^{\mathrm{E}}_v - X^{\mathrm{E}}_{v-1}|$ can be rewritten as $\frac{(n_s-n_1)(n_s-n_2-v+X'_v)}{(n_s-v)(n_s-v+1)}$. Again there are three subcases:

Subcase	Difference	Resulting constraint
$v=1, X'_v=1$	$\frac{(n_s-n_1)(n_s-n_2)}{n_s(n_s-1)}$	$n_1>0 \vee n_2>0$
$v=n_1, X'_v=\max(1,n_1+n_2-n_s)$	$\frac{n_s-n_1-n_2+1}{n_s-n_1+1}$ if $X'_v=1$ 0 if $X'_v=n_1+n_2-n_s$	no constraint
$v=n_1, X'_v=\min(n_1,n_2)$	$\frac{n_s-n_2}{n_s-n_1+1}$ if $X'_v=n_1$ $\frac{n_s-n_1}{n_s-n_1+1}$ if $X'_v=n_2$	no constraint

To conclude, if $n_1+n_2-1\le n_s$, then $|X^{\mathrm{E}}_v - X^{\mathrm{E}}_{v+1}|\le 1$ and Azuma's inequality can be applied.