BOSTON UNIVERSITY
GRADUATE SCHOOL

Dissertation

A COMBINATORIAL NEURAL NETWORK EXHIBITING EPISODIC AND SEMANTIC MEMORY PROPERTIES FOR SPATIOTEMPORAL PATTERNS

By

GERARD J. RINKUS

B.A., University of Rochester, 1983
M.A., Hofstra University, 1986

Submitted in partial fulfillment of the
Requirements for the degree of
Doctor of Philosophy

Approved by

First Reader _____

Daniel H. Bullock, Ph.D.

Associate Professor of Cognitive and Neural Systems

Second Reader _____

Michael A. Cohen, Ph.D.

Associate Professor of Cognitive and Neural Systems

Third Reader _____

Frank Guenther, Ph.D.

Assistant Professor of Cognitive and Neural Systems

Acknowledgments

I thank Prof. Dan Bullock, my major advisor, for his critical analyses of my ideas and for his many suggestions that substantially improved the quality of my theory. I also thank Dan for the guidance that he gave me while I pursued this degree, and for his great example of being a dedicated and responsible scientist, and most of all for encouraging me to blaze my own trail.

I also thank my other readers, Prof. Mike Cohen, Prof. Frank Guenther and Prof. Mike Hasselmo. They invested a great deal of time in reading earlier versions of my thesis and also provided many insightful ideas and challenges that improved my theory and generally furthered my development as a scientist. The interest my advisors had in my work was a great source of inspiration during the development of my theory.

Thanks to my mother who has been waiting for the past four years or so to put that announcement in the Garden City News that her son is a “Doctor”. My mom has always had complete faith in my success even though I had to push the completion date back a few times. Thanks for your love and patience, mom.

Thanks to my father who always told me that anything I can imagine is possible. Dad loved his life and his family to the fullest. Dad was a fair, wise and infinitely patient man who instilled in me the desire to understand the world around me, and to put my heart and full attention into the problem at hand—in other words, to do the job right.

Thanks to my brother, Allan and sister, Sharon, and the rest of my family and friends who have all been extremely supportive and understanding during my pursuit of this degree.

Finally, thanks to my fiancée, Lisa, who came into my life midway during the writing of this thesis and helped me to get through some of the rough spots along the way. Thank you Lisa for the love, happiness and excitement you give me.

Preparation of this report was partially supported by grant ONR N0014-92-J-1309 to Prof. Daniel Bullock and by a research assistantship granted by Steve Grossberg and Gail Carpenter of the Dept. of Cognitive and Neural Systems. I also thank Walter Roy, Patrick Sincebaugh, and Shawn Walsh, of the U.S. Army Research Lab’s Materials Directorate for their support of my research.

To my parents

A COMBINATORIAL NEURAL NETWORK EXHIBITING EPISODIC AND SEMANTIC MEMORY PROPERTIES FOR SPATIOTEMPORAL PATTERNS

(Order Number.)

GERARD J. RINKUS

Boston University Graduate School, 1997
Major Professor: Daniel H. Bullock,
Associate Professor of Cognitive and Neural Systems.

Abstract

A model is described in which three types of memory—episodic memory, complex sequence memory and semantic memory—coexist within a single distributed associative memory. Episodic memory stores traces of specific events. Its basic properties are: high capacity, single-trial learning, memory trace permanence, and ability to store non-orthogonal patterns. Complex sequence memory is the storage of sequences in which states can recur multiple times: e.g. [A B B A C B A]. Semantic memory is general knowledge of the degree of featural overlap between the various objects and events in the world. The model's initial version, TEMECOR-1, exhibits episodic and complex sequence memory properties for both uncorrelated and correlated spatiotemporal patterns. Simulations show that its capacity increases approximately quadratically with the size of the model. An enhanced version of the model, TEMECOR-II, adds semantic memory properties.

The TEMECOR-I model is a two-layer network that uses a sparse, distributed internal representation (IR) scheme in its layer two (L2). Noise and competition allow the IRs of each input state to be chosen in a random fashion. This randomness effects an orthogonalization in the input-to-IR mapping, thereby increasing capacity. Successively activated IRs are linked via Hebbian learning in a matrix of horizontal synapses. Each L2 cell participates in numerous episodic traces. A variable threshold prevents interference between traces during recall.

The random choice of IRs in TEMECOR-I precludes the continuity property of semantic memory: that there be a relationship between the similarity (degree of overlap) of two IRs and the

similarity of the corresponding inputs. To create continuity in TEMECOR-II, the choice of the IR is a function of both noise (Λ) and signals propagating in the L2 horizontal matrix and input-to-IR map. These signals are deterministic and shaped by prior experience. On each time slice, TEMECOR-II computes an expected input based on the history-dependent influences, and then computes the difference between the expected and actual inputs. When the current situation is completely familiar, $\Lambda = 0$ and the choice of IRs is determined by the history-dependent influences. The resulting IR has large overlap with previously used IRs. As perceived novelty increases, so does Λ, with the result that the overlap between the chosen IR and any previously-used IRs decreases.

Contents

List of Tables

List of Figures

Chapter 1. Introduction

An unsupervised, distributed, associative network model of storage, retrieval and recognition of binary spatiotemporal patterns, possessing some of the apparent functional properties of human memory and inspired by some of the more general architectural and dynamical properties of the cortex of the mammalian brain, is proposed. The name of the model, TEMECOR, which stands for *Temporal Episodic Memory using Combinatorial Representations,* reflects its origins as a model of human episodic memory that Tulving (1972) defined as memory for specific events one has experienced. The original model, TEMECOR-I[1], meets several essential requirements of episodic memory--very high capacity, single-trial learning, permanence (i.e., stability) of traces, and the ability to store highly-overlapped spatiotemporal patterns, including *complex state sequences* (CSSs) which are sequences in which the same state can recur multiple times (e.g., [A B B A G C B A D])--however, it fails to possess the crucial property that similar inputs map to similar internal representations—i.e., *continuity*. Therefore the model fails to exhibit similarity-based generalization and categorization, which are the basis of many of those phenomena classed as semantic memory. A second version of the model, TEMECOR-II considerably more complex than the original, adds the property of continuity and therefore constitutes a single associative neural network which exhibits both episodic and semantic memory properties, and which does so for the spatiotemporal pattern domain. TEMECOR-II achieves the continuity property by computing, on each time slice, t, the degree of match, G, between its expected and actual inputs and then adding an amount of noise, inversely proportional to G, into the process of choosing a final internal representation at t. As explained in Sec. 1.5.3, this generally leads to reactivation of old traces (i.e., greater pattern completion) in proportion to the familiarity of inputs, and establishment of new traces (i.e., greater pattern separation) in proportion to the novelty of inputs.

---

[1] A preliminary description of the basic design and principles of TEMECOR-I can be found in Rinkus (1993), although the model has another name in that paper. Rinkus (1995) contains a more complete description TEMECOR-I.

1.1 Semantic Memory

Semantic memory is sometimes simply defined as general knowledge about the world (Squire, 1987). It is that which enables understanding of the meanings of objects and events. According to Tulving (1972), the term “semantic memory” has, from its introduction into the literature [which Tulving ascribes to Quillian (1966)], denoted not merely a static collection of facts about the world but also the means for accessing those facts and for using those facts to solve problems, make logical inferences, and generally, to accomplish the various types of higher-level reasoning tasks humans routinely perform. Typical examples of semantic memory models (Quillian, 1968; Collins & Quillian, 1969; Anderson & Bower, 1973; Collins & Loftus, 1975) consist of a highly structured network of concepts for which logical inferencing procedures can be formally defined. Given the inclusion of reasoning operations (in addition to static facts) in the traditional concept of semantic memory, it is possible to narrowly construe semantic memory-and thus, meaning-as a fundamentally symbolic phenomenon. Indeed, Tulving (1972) summarizes semantic memory as the kind of memory “necessary for the use of language,” and as a “mental thesaurus” containing information about “words and other verbal symbols, their meanings and referents, about relations among them, and about rules, formulas, and algorithms for the manipulation of these symbols, concepts, and relations.” (p. 386)

However, a broader view of semantic memory is espoused herein. In particular, we will consider any piece of higher-order statistical (correlational) information about the world, whether it be linguistically expressible (symbolic) or not (sub-symbolic), as a meaningful “fact” and thus, an item of semantic memory. Thus, the similarity relationships that exist over the set of inputs constitute semantic memory (knowledge). Furthermore, we consider the act of categorization, which accesses such knowledge, to be fully analogous to the operation (process) of traversing an “IS-A” link in a semantic network (Quillian, 1968). It is with these definitions and correspondences in mind that we claim that TEMECOR-II exhibits semantic memory properties. Simulation results demonstrating a) the embedding of similarity relationships in the model's learned mappings between inputs and internal representations, and b) the model's ability to co-categorize similar spatiotemporal events, are given in Sec. 1.3. An additional speculative mechanism for controlling the generality of the information contained in a given retrieval from the model is described in Sec. 4.14.

In this view of the relationship between semantic memory and distributed neural systems, which generally concurs with that expressed in Hinton, McClelland & Rumelhart (1986), many neural network models—e.g., Hopfield (1982), Rumelhart & McClelland (1986), McClelland & Rumelhart (1985), Jordan (1986), Elman (1990), Williams & Zipser (1989), as well as TEMECOR-II are considered to contain semantic memory properties. However, only TEMECOR-II also exhibits the full array of episodic memory properties mentioned above. We will return to a discussion of the relationship between episodic and semantic memory in Sec. 1.3.

1.2 Episodic Memory

1.2.1 Capacity

Episodic or *autobiographical* memories are vivid, perhaps multi-modal, detail-rich memories that can generally last a lifetime even though they are derived from events that occur only once. While it is probably uncontentious to claim that normal human beings have *very* high capacity for storing episodic information, it is intrinsically difficult to quantify (Cohen, 1989, p.120). For example, if the measurement technique relies on verbal reports, then there is an immediate, possibly substantial, loss of information due to passing through the linguistic nexus. Another problem with quantifying episodic information is that it is generally impossible to absolutely determine whether a subject is *confabulating*—i.e., recalling components of distinct episodes as having occurred together. Nevertheless, human episodic capacity does seem to be quite large and there have been many documented cases of people with exceptionally vast episodic memories (Neisser, 1982), for example, Luria's patient “S” (Luria, 1968).

1.2.2 Single-trial learning

It is an open question as to whether a mental experience—i.e., the occurrence of a particular pattern of activation over a region of cortex--that occurs exactly once can last a lifetime. It is most likely the case that very long-lived episodic memories derive from a) actual physical events that were experienced multiple times, or b) mental events (possibly originally derived from external events) that were rehearsed multiple times, or c) a combination of the two. Various studies have shown what is apparent from experience; that recallability of episodic memories increases as a function of subsequent rehearsal (i.e. reminiscence) (Rubin & Kozin, 1984). Nevertheless, the

TEMECOR models[2], as described herein, are capable of true single-trial learning—i.e., neither multiple overt trials nor rehearsal are needed--and, thus exhibit a competence that most likely exceeds that of average human beings. However, TEMECOR also has two problems as a model of human memory, one from the neurobiological standpoint and one from the psychological standpoint. These problems will be discussed later. At this point, we simply want to note that the proposed additions to TEMECOR-II-specifically, the addition of a hippocampal analog (see Sec. 4.13), required to remove these two problems, also reduces this unrealistic single-trial capability back to a level more commensurate with that of normal humans, in which either multiple presentations or rehearsal are necessary to permanently embed episodic memories.

1.2.3 Stability

One of the preeminent properties of episodic memories is that they can last for an entire human lifetime; they are extremely stable. Furthermore, individual traces can remain unaccessed (at least consciously) for many years, during which other traces are accessed frequently, and then suddenly be called to mind by some fortuitous arrangement of stimuli. TEMECOR exhibits this type of stability. This is due to the fact that the modifiable weights in the model, which are {0,1}-valued, can only increase. Information can only be lost because of interference in TEMECOR. Interference increases as a function of *saturation*—i.e., the proportion of weights that have been increased. Saturation increases as a function of the number of unique inputs stored, but *not* as a function of their order of presentation, nor of the frequency of presentation of each input.

This raises the question as to how saturation is prevented in the model. As more and more patterns are presented to the model, more and more synaptic weights are increased to the maximum weight of one. If all synaptic weights are increased, then all information is lost. Two distinct mechanisms for reducing the rate of saturation are discussed in Sec. 4.9.

Memory schemes employing Backpropagation, on the other hand, can lose memories due to the repeated presentation of a single or small group of patterns--i.e., *catastrophic interference* (McCloskey & Cohen, 1989). That is, memory loss depends not only on the number of unique patterns presented but also on the frequency and order of presentation of those patterns. In particular, under the sequential training paradigm of McCloskey & Cohen (1989), the model they

[2] We use the convention throughout that when the statement being made applies to both TEMECOR-I and TEMECOR-II, we simply say TEMECOR.

tested showed almost complete forgetting (i.e., 90%) of the first list of exemplars within the first few training trials of the second list (even under the easiest paradigm for deciding correct vs. incorrect responses). Such memory schemes are unstable regardless of how saturated—i.e., how close to capacity—they are. If the goal is to model episodic memory, which by its nature records statistically rare events, then we should prefer models for which stability depends only on degree of saturation, not on the frequency and order of presentation of exemplars.

This issue of stability constitutes one of the principal motivations for the *Adaptive Resonance Theory* (ART) developed in Grossberg (1976, 1978), Carpenter & Grossberg (1987). Grossberg describes the problem as the *stability/plasticity dilemma*: ideally, a system should remain capable of learning when important new inputs occur but it must also prevent important old traces from being overwritten. The means by which the ART models achieve this property is by computing the degree of match between the expected input and the actual input. A general discussion of the stability issue and of the relationship between ART, TEMECOR and Backpropagation is provided in Sec. 1.6

1.2.4 Non-orthogonal patterns

Another important characteristic of episodic memory is that there may generally be a great deal of featural overlap over the set of individual episodes comprising a given person's episodic memory. For example, one may be able to recall literally hundreds of episodes involving his father. Here, we are allowing that features may be quite high-level, i.e., father. As with capacity, quantification of the amount of overlap over the set of episodes comprising a given person's episodic memory is intrinsically difficult.

The data sets used in the simulations in this thesis can be divided into two categories, uncorrelated and correlated. The uncorrelated sets contain significant overlap between episodes and the correlated sets, even more. Simulations reported herein typically had an input layer consisting of 100 binary feature detectors. In the uncorrelated case, typically S = 20 of the M = 100 features were chosen at random to be active on any given time slice of an episode. Thus, if P episodes, each having T time slices, have been presented, then any given feature is expected to have occurred $(P \times T \times S)/M$ times. Thus, in the largest simulations (see Table 3.6), features occurred an average of over 6,000 times each.

The following method was used to generate the correlated pattern sets. First, an *alphabet* (i.e., *symbol set*) of *U* unique states, each consisting of *S* (out of *M*) active features, was built. The time slices comprising the episodes were then randomly chosen (with replacement) from this alphabet of states. Thus, formally, these data sets are sets of complex state sequences (CSSs). Assuming P episodes having *T* time slices each, the expected number of occurrences of a state is $(P \times T)/U$. The largest correlated pattern simulations (i.e., involving complex sequences) had *U = 100* unique states and about 2670 episodes, each consisting of *T = 10* states, were learned to criterion. Thus, this simulation involved a total of 26,700 state instances over an alphabet of only 100 states, yielding an average of about 267 instances of each state. At the time of this writing, I am aware of no other report in which a set of CSSs of this size and complexity has been successfully learned.

1.3 Relationship of Episodic and Semantic Memory

We have described semantic memory essentially as knowledge of the higher-order statistics of the input set (general information) and episodic memory as knowledge of the specific details of individual exemplars comprising the input set (specific information). The nature of the relationship between these two different types of information remains a major open question of cognitive psychology and of cognitive neuroscience. The phenomenon known as *confabulation*, in which a person will erroneously recall components from various distinct episodes as having occurred together provides one strong indication of the interaction between episodic and semantic memory. Studies by Loftus (1977) show that when people confabulate, they make substitutions that are semantically feasible. For example, a person might remember meeting someone at a train station 50 years earlier, when in fact it was a bus station, but it is generally far less likely that he will remember having met the person in a bedroom, or in a forest, or, to carry the point to the extreme, in a refrigerator. The likelihoods of substituting “the train station”, “the bedroom”, “the forest” or “the refrigerator” for the “bus station” correlate with their respective semantic distances from (i.e. featural similarities/dissimilarities with) “bus station”.

Perhaps the most salient fact regarding the relationship between episodic and semantic memory is that all information enters the human system one exemplar at a time. Nevertheless, humans naturally notice similarities and dissimilarities and form generalizations and categories. This has been demonstrated for the spatial pattern domain in *schema abstraction* studies of Posner & Keele (1968) and Homa, Cross, Cornell, Goldman & Schwartz (1973). The implicit learning of grammar

studies of Reber (1967) demonstrate this for the spatiotemporal domain. The most striking fact shown in some of these studies is that after experiencing some number of exemplars of a given category, subjects were better at recognizing the prototype—i.e., central tendency—of the category than the individual exemplars even though the prototype was never experienced. The fact that this effect generally increases over time (Homa, et al., 1973, Posner & Keele, 1970) was taken to imply the existence of separate internal representations that embodied general information which, for example, underlie performance of generalization and categorization tasks (Medin & Schaffer, 1978; Brooks, 1987).

However, there is much recent psychological evidence (Brooks, 1978, 1987; Whittlesea, 1987, 1989; Vokey & Brooks, 1992; Whittlesea & Dorken, 1993) supporting the view that one's general knowledge of the correlational and categorical structure of the world may actually be distributed amongst the set of memory traces corresponding to the individual exemplars that have been experienced and that "performance in conceptual and perceptual tasks....appears to be determined by memory for particular events" (Whittlesea, 1989, p.78). Whittlesea describes the structure and findings typical of the studies cited above as follows.

> *"...I constructed a stimulus domain of letter-string stimuli in which the similarity of test items to the prototype could be manipulated independently of their similarity to particular training instances (Whittlesea, 1987). Subjects were required to copy selected training stimuli, and then to identify briefly presented test items. The accuracy of identification was found to be correlated with the similarity of probes to the set of training items, but not systematically related to the typicality of the probes. This denies the conclusion that regular or typical stimulus properties automatically have a special status in memory, and suggests that performance in unreflective tasks such as perception of category members may also rely on idiosyncratic information about particular events."*

In the terminology of Vokey & Brooks (1992), the *distributive* view which assumes general knowledge is implicit in the set of episodic traces contrasts with the *abstractive* view which assumes that explicit, centralized, canonical representations of categories and correlations are constructed during the learning phase in which the individual exemplars are presented.

Theories accounting for both episodic and semantic memory can be divided into three classes, the first of which corresponds most closely to Brooks' *abstractive* models, and the latter two of which correspond to Brooks' *distributive* models.

1. Episodic memory (EM) and semantic memory (SM) are physically disjoint stores. Of course, they interact with each other, but episodic traces are physically disjoint from the semantic traces.
2. There is only EM. The only memory traces actually stored in the system are those corresponding to individual episodes. Brooks (1987) refers to such theories as *post-computational*, because the processes that compute the correlational information implicit in the set of episodic traces occur at retrieval time that is necessarily subsequent to initial encoding. The *multiple-trace* theory, MINERVA-2, of Hintzman (1986) and the *context model* of Smith & Medin (1981) are such theories.
3. EM and SM are physically overlapped. The same physical substrate is used for both episodic and semantic information. Individual episodes are all that is ever explicitly stored however general knowledge of the correlations in that set of episodes is present in the particular structure of the overlapped traces. The changes to memory incurred in storing a new episode automatically cause changes to the general knowledge contained in the memory. McClelland & Rumelhart (1985) propose such a model that they describe as a "distributed, superpositional approach to memory" (p.160). These authors point out that although "generalizations emerge from the superposition of specific memory traces" in both their model and post-computational models, that superposition occurs at time of learning in their model whereas, as stated above, it occurs at time of retrieval in the post-computational models. TEMECOR-II is also an instance of this class of model.

The traditional semantic memory models of (Quillian, 1968; Collins & Quillian, 1969; Anderson & Bower, 1973; Collins & Loftus, 1975) as well as the more recent models of Rumelhart & Norman (1978) and Schank (1982) implicitly fall into the first class because they assume, from the outset, generic (and localist) representations of concepts. Such semantic memories could potentially be linked to episodic stores, however, as Hintzman (1986, p. 423) points out, such theories encounter difficulty in explaining "how the abstract knowledge that is assumed to be

stored explicitly in semantic memory was learned originally and how it is modified by experience." Indeed, this is at the core of the debate between symbolic (artificial intelligence) and subsymbolic (connectionist, neural network) approaches.

Although theories of the class b (above) address the issue of how abstract knowledge is distilled from the individual episodes (which are the direct objects of experience) and can explain a large number of experimental findings, they have two problems. First, such models are extremely inefficient in terms of storage because they assume that that every episode, no matter how similar it is to previous episodes, gets its own trace that is physically disjoint from all others. The second potential problem is that as the number of stored episodes gets large, the temporal duration of the computations extracting general knowledge, which are assumed to take place at retrieval time, will grow.

Distributed neural network models are natural candidates for the third class, however, thus far, there have been very few demonstrations of neural network models exhibiting both semantic and episodic memory.[3] One such model, that of McClelland & Rumelhart (1985), is shown to be capable of storing both general and specific information, however it has a number of shortcomings as a model of episodic and semantic memory. Their simulation demonstrating simultaneous storage of general and specific knowledge is somewhat strained in two respects. It requires that the prototype pattern is itself presented, multiple times, as an input. Thus, in their simulation, the "general" information is explicitly presented rather than having to be derived by the model itself. Further, as the authors clearly point out, it requires that the specific items to be stored are presented as often as the prototype pattern. This is clearly contrary to the single-trial property of episodic encoding. Moreover, the fact that it requires many presentations of each input (in contrast to the additional constraint that the specific items be presented as often as the general items) is itself a violation of the single-trial learning criterion for episodic memory. Finally, it is a supervised model that uses the Delta learning rule and thus can only learn linear mappings.

Despite its shortcomings, this earlier model of McClelland & Rumelhart (1985) is distinctive because it is *monolithic*. That is, it is not a multi-component model where various components

---

[3] There has been a similar dearth of exploration of the episodic-semantic relationship within the neurophysiological community. Squire (1987, p. 172) states that thus far there has been "almost no effort directed toward this problem". Lynch & Granger (1994, p. 66) write, "It is indeed curious that rats are seldom tested on problems involving stable encoding, rapid acquisition, and very large numbers of similar and specified cues."

accomplish various functions that are differentially related to episodic and semantic memory. Rather, it is one monolithic, homogeneous and simple architecture utilizing one simple learning rule. We raise this issue because this criterion can be used to distinguish TEMECOR-II, which is also monolithic, from another non-monolithic class of models that potentially address both episodic and semantic memory. Specifically, this is the class of models that contain a "neocortical" component and a "hippocampal" component, two examples of which are McClelland, McNaughton & O'Reilly (1994) and Murre (1995). While neither of these two papers focuses on the issue of episodic vs. semantic memory, they do discuss the issue and we will discuss these models in Sec. 2.3. Although the current, monolithic TEMECOR-II achieves semantic, episodic and sequence memory properties, it is clear from the experimental and clinical literatures that the hippocampal complex is of fundamental importance to memory. In fact, TEMECOR-II has two significant problems as a model of human memory--one from the neurobiological standpoint and one from the psychological standpoint--that can be solved by inclusion of a hippocampal component. A speculative outline for adding this component is given in Sec. 4.13.

1.4 Complex State Sequence (CSS) Memory

As indicated in the previous section, the model is capable of remembering large sets of CSSs. This problem lies at the heart of speech, and more generally, language processing. This is because, to first approximation, linguistic objects are formally complex state sequences, and normal humans have extremely large capacities for storing such objects. For example, all spoken English words are sequences from an alphabet of about 40 phonemes (states). Similarly, all English sentences are sequences from an alphabet of many tens of thousands of words. Oldfield (1963) estimated that the average young university-educated person knows the meaning of about 75,000 words. This corresponds to as many as about 90,000 distinguishable word forms[4] in such a person's memory. If we assume that on average, words contain five phonemes, then the average number of instances of any phoneme is 9,000. This suggests a very large *branching factor* and thus a highly complex set of sequences.

Recently there has been a great deal of research on recurrent backpropagation (RBP) models (Jordan, 1986; Elman, 1990; Williams & Zipser, 1989). For the most part, these models have been

[4] Many root words have many different variants (e.g. "jump", "jumped", "jumper", etc.) that also have different meanings and thus must correspond to separate entries in one's lexicon.

applied to the problem of learning to recognize state sequences generated by finite-state automata (FSAs). This is a pattern recognition task and thus essentially requires learning the correlational structure of the input set. This research is of particular interest in this thesis because FSAs generate CSSs. While these models have been very successful at the recognition (or prediction) task—i.e., at extracting general knowledge of the input domain, they have not been shown to be able to recall the inputs episodically. In fact, the clear implication of the research so far (Cleeremans, 1993; Hochreiter & Schmidhuber, 1995) is that it is unlikely that the RBP models can be endowed with episodic memory capability. We review these models in Sec. 2.2.3.

The ability to process complex sequences has also been a central focus of the hippocampal model developed by Levy and his colleagues (Levy, 1989; Minai & Levy, 1993; Minai, Barrows & Levy, 1994; Levy, Wu & Baxter, 1995; Levy & Wu, 1995; Wu & Levy, 1995). We will review this work in Sec. 2.2.4.

1.5 Summary of TEMECOR

The summary of the model is broken into several sections. The first section describes the underlying representational principle that is common to both versions of the model. Following that, the basic architecture and operation of TEMECOR-I are summarized. Then the next section describes the basic principle by which continuity is added to the model. Finally, a brief summary of TEMECOR-II, which has significant architectural and operational differences from TEMECOR-I is given.

1.5.1 The Underlying Representational Principle

The large capacities exhibited by TEMECOR-I, and to a lesser extent, by TEMECOR-II, derive from the use of a *combinatorial* representation scheme that is clearly described in Willshaw, Buneman & Longuet-Higgins (1969). The basic principle is illustrated in Figure 1.1. Panel a depicts a simple network in which an A pattern, A_1, is associated to a B pattern, B_1, by setting all connections from active A_1 cells to active B_1 cells to one. Panel b depicts the association of another, partially overlapping pair, A_2-B_2. Panel c shows what happens when, following the two learning trials, we reinstate A_1. In particular, each cell in B_1 will receive a total of three large active inputs, whereas the *spurious* B cells—i.e., those in B_2--will receive only one large active input. Similarly, panel d shows that if we reinstate A_2, the cells in B_2 receive three large active inputs

whereas the spurious cells receive only one. Thus, to achieve the selective reactivation of only the correct B-cells, we can impose a constraint whereby only B-cells with total inputs meeting or exceeding some threshold, which in this case could be either two or three, can become active. This is the essential insight underlying the *Correlograph* model of Willshaw et al. (1969) that has been shown to yield very high capacities, especially in the *sparse coding limit*—i.e., where the ratio of the number of active cells in a layer to the total number of cells in the layer (i.e. the *coding rate*) is very small. In particular, assuming both layers have n cells and both A and B patterns have m active cells, then Willshaw's analysis [via McClelland (1986)] showed that the number of patterns, *r*, that could be stored before the probability of having at least one spurious B-cell is one, is:

$$r \leq 0.69(n/m)^2$$

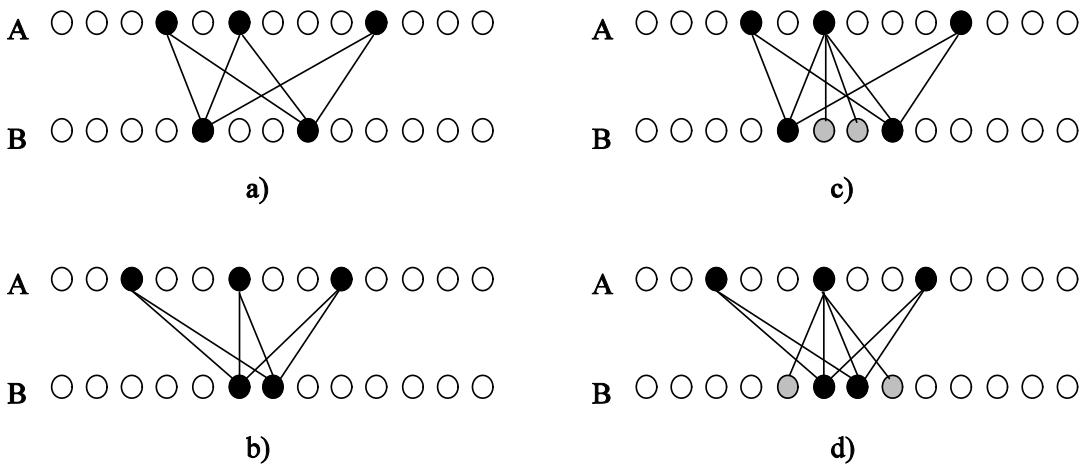

Figure 1.1: Depiction of the essential idea underlying the use of combinatorial representations as described for the Correlograph model of Willshaw et al. (1969). Panels a and b show the learning of two partially overlapping associations, A_1-B_1 and A_2-B_2. Panels c and d show that both associations can be recovered perfectly if B-cells are only allowed to become active if their total input meets or exceeds a threshold which in this case could be either two or three.

The TEMECOR model is based on this same fundamental principle. However it adds another level of complexity to the model. In particular, rather than being simple, undifferentiated fields of cells, the representational fields of the model are organized in *competitive modules* (CMs), wherein exactly one cell can become active on any given time slice. Although the *winner-take-all* (WTA)

dynamics of these CMs are not explicitly modeled herein, they could be implemented, for example, in terms of the recurrent competitive field theory presented in Grossberg (1973).

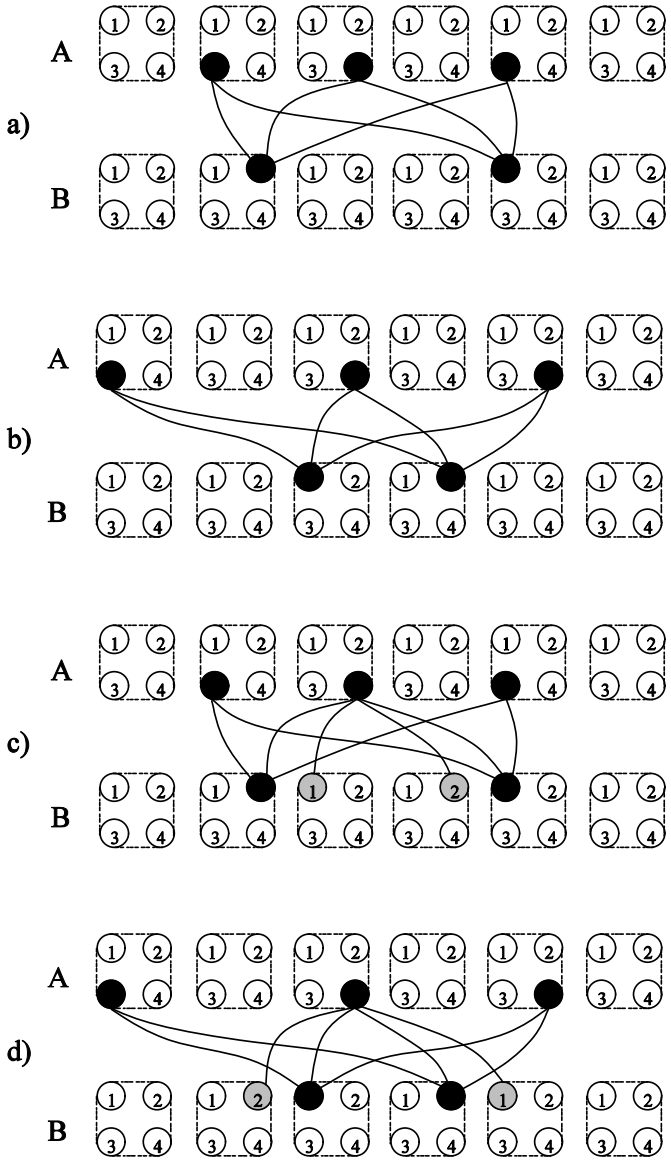

Figure 1.2: Illustration of basic combinatorial representation principle in conjunction with the use of winner-take-all competitive modules (CMs). The same descriptive remarks, for each panel, given for the previous figure apply here as well.

Figure 1.2 illustrates the same principle as did Figure 1.1 except that it incorporates CMs. Once again, the constraint that B-cells can only become active if they have sufficient total input allows selective reactivation of the correct associations despite overlap in the mappings. The capacity analysis is unaffected by the grouping of the fields into CMs-only the coding rates matter, thus Eq. 1.1 applies for the case of CMs as well. The partitioning of TEMECOR's principle representational field, its layer 2 (L2), into CMs was done for two reasons.

1. TEMECOR is envisioned as a general model of cortex, particularly deeper cortices like entorhinal cortex. The CM is considered to be analogous to the cortical *mini-column* (Szentagothai, 1975; Mountcastle, 1978; Eccles, 1981), a group of about 100 excitatory pyramidal cells together with 1-2 inhibitory cells, which has been found to be a rather ubiquitous feature of neocortex, including piriform/entorhinal cortex (Van Hoesen & Pandya, 1975). Given the lack of direct neurophysiological evidence that cortical mini-columns function in a winner-take-all fashion, the analogy between CMs and mini-columns is a speculative hypothesis, however other modelers make a similar association (Coultrip & Granger, 1994).
2. It provides a principled means of ensuring a small coding rate at L2, which, as per Eq. 1.1, yields the large capacity demonstrated in the simulations reported herein. That is, if there are Q CMs each having K cells, then the maximum possible coding rate (in the case where we assume all CMs are active in every representation) is 1/K, e.g. 1/100 (using the estimate of about 100 excitatory pyramidal cells per mini-column).

1.5.2 Basic version: TEMECOR-I

Figures 1.1 and 1.2 describe the basic distributed, combinatorial representational principle in the context of purely spatial mappings. Figure 1.3 summarizes the full, two-layer architecture of TEMECOR-I, which is capable of remembering numerous spatiotemporal binary feature patterns presented from the environment at its input layer, layer 1 (L1). The excitatory, but non-plastic matrix of connections from L1 to L2 is called the *feedforward* projection (F-projection).[5] Note that the binary-valued feature-detecting cells of L1 are connected 1-to-1 with the CMs of L2. The

[5] TEMECOR-II's F-projection differs from that of TEMECOR-I, both topologically and in the fact that it is plastic.

corresponding reverse or *reciprocal* matrix is called the R-projection. Finally, there is an intra-L2, *horizontal* matrix, the H-projection, which interconnects, nearly fully, the L2 cells.

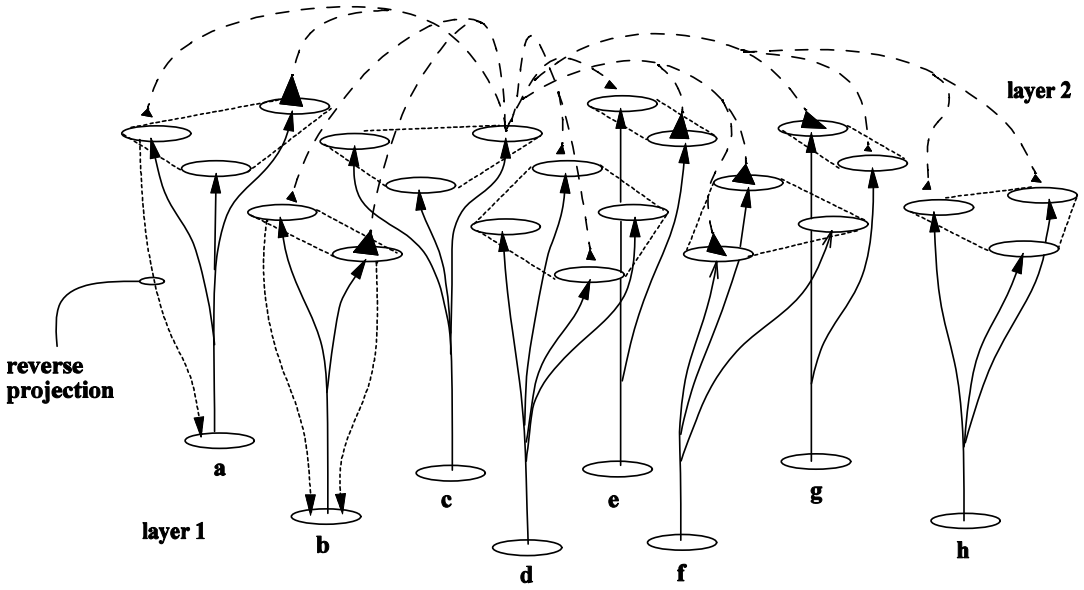

Figure 1.3: TEMECOR-I has two layers. Some of the horizontal connections emanating from one L2 cell are depicted with dashed lines ending in either large (weight = 1) or small (weight = 0) black synapses. Only a few sample reverse (i.e., top-down) projections are shown. This figure will be repeated in Ch. 3 and discussed in more detail at that time.

TEMECOR-I's "life" consists of a learning phase followed by a recall phase (i.e., performance phase).[6] During learning, spatiotemporal patterns are presented, once each, to the model. On each time slice of each episode, a spatial input pattern—or, *L1 code*—is presented, and a corresponding internal representation—or, *L2 code*—is chosen. This L2 code is then linked via Hebbian learning in the H-projection to the L2 code chosen on the next time slice. This continues for the duration of the entire episode, resulting in the formation of a spatiotemporal memory trace of the episode in L2.

Figure 1.4 depicts the sequence of steps underlying the embedding of a memory trace in the H-projection. Panel a depicts the activation of the L1 cells corresponding to the features present on

[6] In contrast, one of the important properties of TEMECOR-II is that it does *not* require the artificial division of its existence into a learning phase and a recall phase. This point will be explained shortly.

the first time slice of some episode. We refer to this as step 1 of time slice, t = 1. Panel b depicts step 2 of t = 1 in which an L2 code (internal representation) has been chosen. An L2 code consists of one L2 cell chosen as winner in each *active* CM--that is, each CM whose corresponding L1 cell is active. The winners are chosen at random within their respective CMs. Note that since the F- and R-projections are non-plastic there is no adaptive linking between the L1 and L2 codes. The only linking (learning) done in TEMECOR-I is between successively active L2 codes. Panel c depicts stage 1 of t = 2 in which a different L1 code (input) has become active. It also shows the fading activations of the previous L2 code (gray cells). Panel d depicts stage 2 of t = 2 in which a corresponding L2 code has been chosen. Finally, panel d also shows the H-synapses that would be increased in this case (dotted lines).

It is important to note that TEMECOR's architecture is *not* partitioned into distinct fields of cells dedicated to representing distinct time slices of input within some temporal window, as for example is the case in the *Time Delay Neural Network* (TDNN) model of Waibel (1989). In particular, both L1 cells (features) and L2 cells can be active on multiple consecutive time slices. This point is emphasized here because subsequent figures used to explain the theory often involve spatiotemporal patterns in which features occur on only one time slice. This is done only to keep the figures readable.

Operation during the recall phase is as follows. In order to test recall of an episode, that episode's first time slice's L2 code is reinstated. This episode-initial L2 code then causes the next L2 code of the episode to be reinstated and so on, until the last time slice of the trace has read-out. The threshold-based mechanism described in Figure 1.2 ensures that selective reactivation of the correct L2 codes occurs on each time slice. In addition, the L2 code that becomes active on each time slice sends signals, via the R-projection, to cause the associated L1 code to become active as well. Figure 1.5 depicts the sequence of steps underlying the read-out of a previously stored episodic trace. In panel a, the episode-initial L2 code has been reinstated. The use of L2 codes as prompts rather than L1 codes is an unrealistic feature of TEMECOR-I. In reality, prompts come from the environment that interacts directly with L1, not L2. However, this unrealistic feature has little bearing on TEMECOR-I's primary result--i.e., faster-than-linear capacity scaling in the number of cells in the model. More importantly, this problem is removed in TEMECOR-II in which there is adaptive linking between the L1 and L2 codes and L1 prompts are used. In panel b, signals are traversing both the H- and R-projections. We assume that the R-signals reach the L1

cells and cause them to become active on the same time slice, whereas the H-signals take one time slice to propagate and cause the next L2 code to become active on the next time slice, as shown in panel c; note the fading activation of the previously active cells (shown in gray). Finally, the L2 code at t = 2 causes its associated L1 code to become active. No further activation takes place because the particular set of cells comprising the t = 2 L2 code have not been linked to any other L2 code.

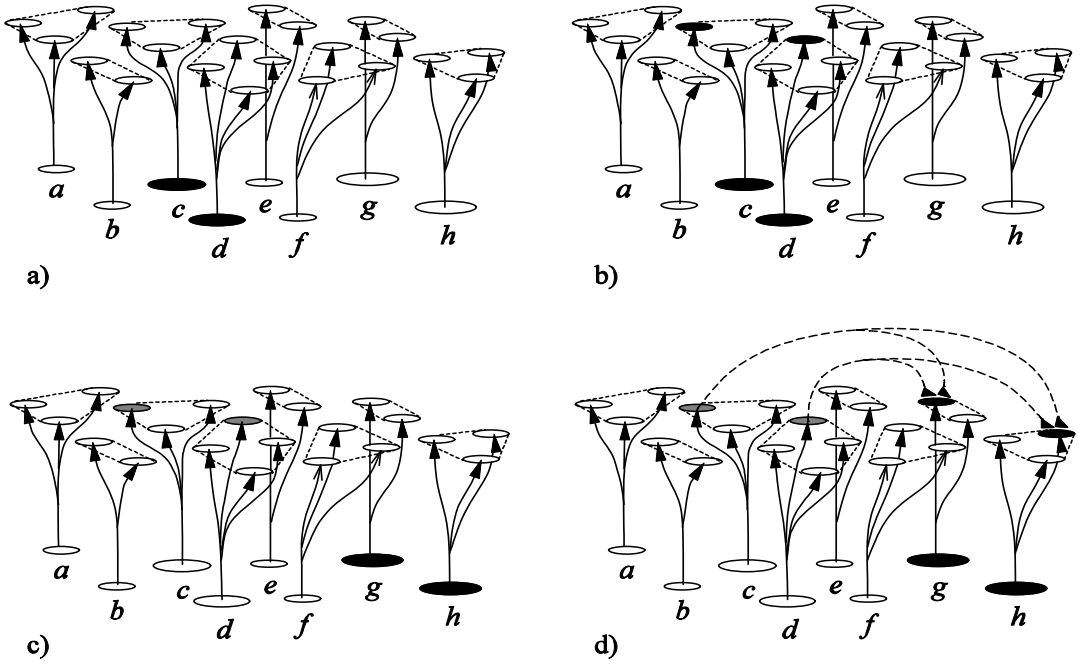

Figure 1.4: Panel a depicts the activation of the L1 cells corresponding to the features present on the first time slice (t = 1) of some episode. Panel b depicts the L2 code chosen to represent the L1 code. In panel c, a new L1 code is active. Panel d shows the L2 code chosen to represent the new input and the resultant learning in the H-projection.

This strategy of choosing winners in active CMs completely at random, on learning trials, has two major implications.

1. It leads to maximal separation over the set of chosen internal representations, on average, and thus to maximal capacity.
2. It precludes the learned mapping of the H-projection from having the property that similar L2 codes lead to similar successor L2 codes. That is, the H-mapping does not have the

property of *continuity*, and thus does not allow similarity-based generalization and categorization.

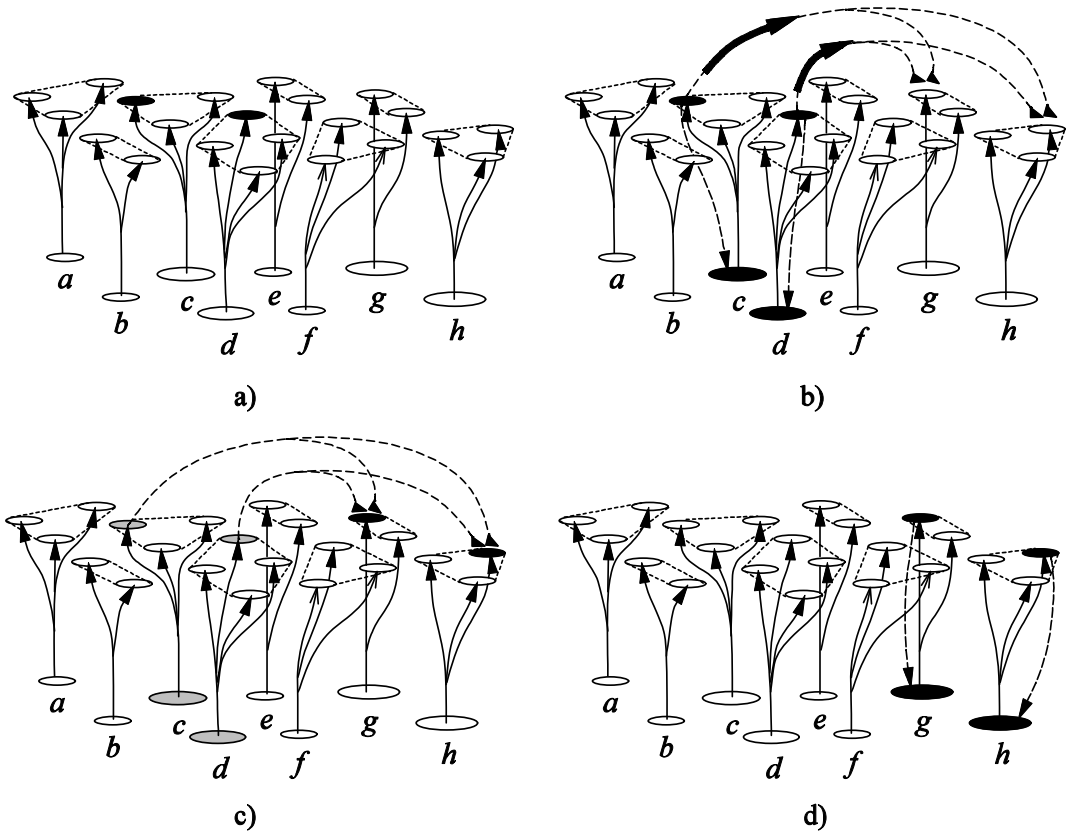

Figure 1.5: Panel a depicts the reinstatement of an episode-initial L2 code. See text for discussion of this step. Panel b depicts the reinstatement of the associated L1 code via signals propagating in the R-projection. Panel c depicts reinstatement of the next L2 code based on the signals arriving via the H-projection. Panel d depicts reinstatement of the corresponding L1 code.

The fact that, during learning, winners are chosen completely at random in TEMECOR-I is tantamount to a complete lack of dependence of the choice of winners on any model quantities. In particular, the signals propagating in the H-projection--hereinafter, the *H-vector*, which reflect prior learning, have no influence in determining the winners. However, as Figure 1.5 suggests, transmission of H-signals occurs at full strength during recall trials and the H-vector fully determines the winners. Thus the model has two different dynamics, one for learning, one for recall. We can model the winner selection process as depending on a mixture of two influences, noise and signals propagating in the plastic associative mappings (i.e., the H-projection in the case

of TEMECOR-I). In this case, the distinction between the two modes is simply that in learning mode, noise is very high and "drowns out" the deterministic, learned signals, whereas in recall mode, noise is zero, thus allowing the deterministic signals to fully determine the winners.[7] The question naturally arises: can anything be gained by varying, in a more graded fashion, the relative amount of noise in the winner selection process? The answer is "yes" and exploration of this issue has led to a method for adding continuity to the learned mappings of TEMECOR-I, resulting in TEMECOR-II. In particular, in TEMECOR-II, the relative amount of noise added depends on the computed degree of similarity (match) between TEMECOR-II's expected input at t and the actual input at t. The basic principle is explained in the next section.

1.5.3 Continuity via Match-contingent Noise

Continuity results if the internal representation that would be chosen for an input based solely on the deterministic, history-dependent signals that arise when the input is presented, is randomly changed by an amount that is inversely proportional to the similarity of that input and the set of previously-experienced inputs. This is explained in terms of the generic, spatial, associative memory model of Figure 1.6 in which A patterns (i.e., inputs) are mapped to B patterns [internal representations (IRs)]. Suppose that the mapping between an input, A^1, and an IR, B^1, depicted in panel a, has been learned previously. The solid lines connecting A^1 to B^1 denote the increased connections (weights). Panel b shows another input, A^2, having substantial overlap--i.e., similarity--with A^1, which results in strong, albeit sub-maximal, input to the cells comprising B^1 (shaded dark gray to reflect this level of support). We will refer to the set of IR cells receiving the highest amount of input as the *most-highly-implicated* IR. Now suppose that due to the sub-maximal level of support--i.e., a high but sub-maximal match, a small amount of noise is added into the final selection of cells to become active at layer B, resulting in (panel c) a final IR, B^2, slightly different from, although still substantially overlapping, B^1; specifically, $\left|B^1 \cap B^2\right| = 2$. The new learning that would occur in this case is depicted with dashed lines in panel c. Panel d shows another input, A^3,

[7] One could equivalently assume a certain baseline level of noise in the cortex and imagine that it is the relative strength of the deterministic signals that are directly modulated. This is probably closer to reality and in particular maps more directly onto the proposal by Hasselmo (1994, 1995) of a mechanism for setting the global dynamics (i.e., learning vs. recall) based on modulation of acetylcholine (ACh) levels. The relation of this work to TEMECOR will be addressed in more detail in Ch. 4.

having a smaller overlap with A^1, reflected in the light gray shading of cells comprising B^1. Since the similarity between A^3 and A^1 is less than that between A^2 and A^1, relatively more noise is added into the process of choosing the IR, yielding a B^3 having smaller overlap with B^1 than does B^2; i.e., $\left|B^1 \cap B^3\right| = 1$. This example shows the general trend that increased similarity between current input and previous inputs, causes increased similarity between the trace of the current input and pre-existing traces; i.e. continuity.

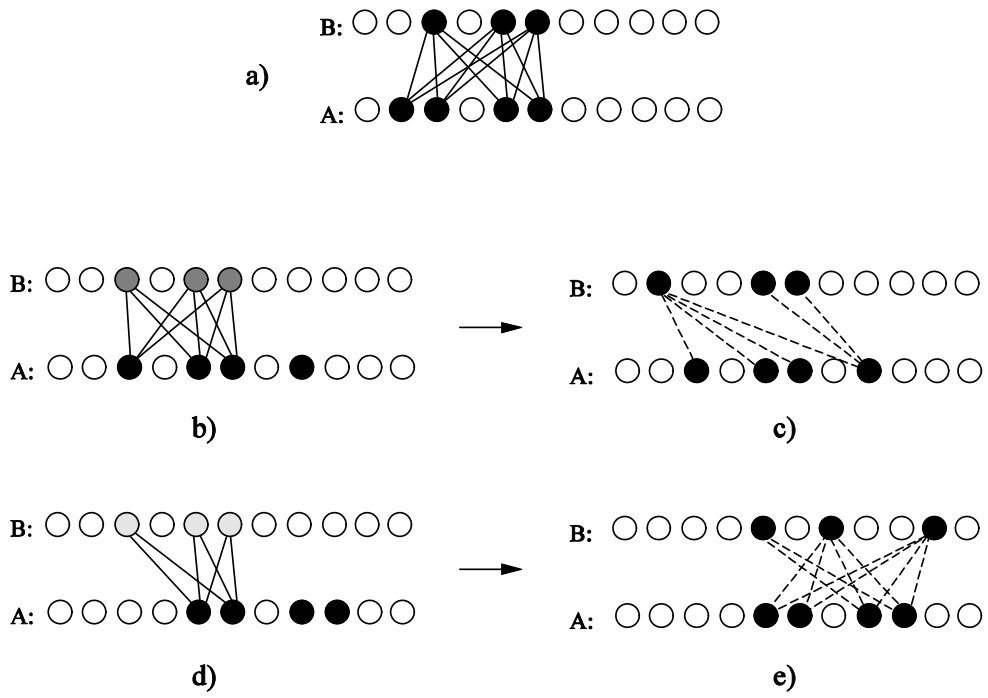

Figure 1.6: The basic principle, used in TEMECOR-II, whereby addition of an amount of noise, inversely proportional to the similarity of current input to the set of previously learned inputs, results in a final mapping having the property of continuity. a) A pre-existing learned mapping between A^1 and B^1. b) Another input, A^2, highly similar to A^1. A relatively small amount of noise is added into the final choice of B^2 which thus, has high overlap with B^1, as seen in panel c. d) Another input, A^3, that is much less similar to A^1. A larger amount of noise is added into the winner selection process, resulting in a B^3 having smaller overlap with B^1 than does B^2. See text for more explanation.

In the limiting case in which the current input has no overlap with—i.e., no similarity to—the previous inputs, the IR choice process becomes a completely random process, resulting in the minimal expected overlap between the resulting IR and the set of pre-existing IRs. In the opposite limiting case in which the current input is identical to some previous input, zero noise is added. Thus, the IR choice process becomes completely deterministic, resulting in the reactivation of the IR corresponding to the matching previous input. We refer to the event in which a final winner--i.e., one resulting after noise has been added--is not one of the most-highly-implicated cells as an instance of *winner-flip*.

1.5.4 Enhanced version: TEMECOR-II

The two most important goals driving the development of TEMECOR-II were: a) adding the capability to use L1 prompts instead of L2 prompts, and b) the addition of continuity. Attainment of these goals required several architectural modifications that are summarized in Figure 1.7. Specifically, the changes are:

a) There is no longer a 1-to-1 correspondence between L1 cells and L2 CMs. All L1 cells contact all L2 cells via the F-projection and vice versa, via the R-projection.
b) The F- and R-projections are plastic.
c) Some additional circuitry that computes, on each time slice, the match between the expected and actual inputs. Most of this circuitry is local to the CM and is not shown in Figure 1.7, however, as the figure suggests, some is distinct from local CM circuitry. In fact, it may be possible to remove all non-local circuitry (i.e., computations) from the model, but this is a subject of future research.

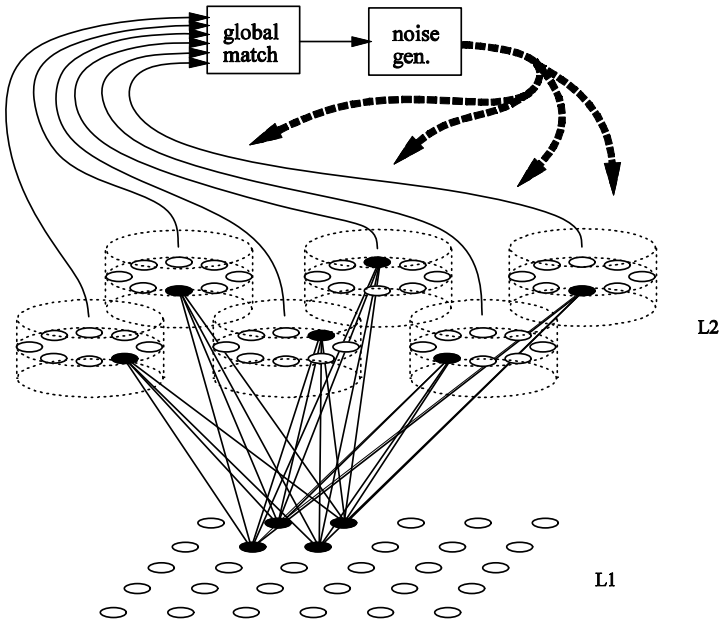

Figure 1.7: The whole TEMECOR-II model. The cylinders are intended to suggest the mini-columns of cortex. The lines leading to the global matching module carry the results of the local matching computations, explained in Ch. 4, that take place within each CM. The global degree of match is then used to determine how much noise to inject into the winner selection process. Note the horizontal connections are not depicted.

The various properties relevant to episodic, semantic and sequence memory, which TEMECOR-II exhibits are summarized in the following list. These properties are *in addition* to those exhibited by TEMECOR-I.

a) It exhibits similarity-based generalization and categorization in the spatiotemporal domain. These properties are evidenced in the simulation results of Ch. 4.
b) It does not require the artificial division of its "life" into a learning phase and a recall phase. The degree of match between expected and actual input on a given time slice effectively and automatically determines (in a probabilistic sense) the extent of learning that will obtain on that time slice. The general principle that the more novel the L2 code is, the more opportunity for synaptic increases there are can be seen in Figure 4.1. The number of newly increased synapses (dashed lines in panels c and d) is higher for the more novel pattern, A^3. Since the novelty of the L2 code is directly influenced by the amount of noise added to the winner selection process,

it follows that the rate of learning is also automatically modulated by the degree of noise added to the winner selection process. Other mechanisms for modulating the rate of learning can be used in conjunction with this indirect mechanism. In particular, if the model is generalized to have continuous weights, then a learning rate parameter could be incorporated to the learning rules of the model, thus providing a more direct control of the amount of learning that obtains on a given time slice. As will be shown, the model's dynamics can shift along the learning-recall continuum within a single episode.

c) It performs completion of spatiotemporal patterns given episode-initial prompts, but also given prompts that originally occurred at a mid-sequence position as long as the prompt is unambiguous. Thus, if the model has previously-experienced the sequence, [A B C D E], then if prompted with C, it will read out [D E].

d) When given an ambiguous prompt, multiple competing expectations (i.e., hypotheses) become active in the system. As subsequent disambiguating information (i.e., successive states of the prompt) enter the system, the unmatched expectations (i.e., disconfirmed hypotheses) fade away.

1.6 Code Stability and Expectation Match/Mismatch

Code stability--i.e., permanence of memory traces--is a central issue in the design of adaptive systems. In fact it is a principal motivating factor in the development of *Adaptive Resonance Theory* (ART) (Grossberg, 1976; Carpenter & Grossberg, 1987). Grossberg (1980, 1982) has identified this as the *stability/plasticity dilemma*: ideally, a system should remain capable of learning when important new inputs occur but it must also prevent important old traces from being overwritten, even if unaccessed for very long periods. ART achieves this property by introducing a separate subsystem-the *orienting* system--that, in conjunction with the attentional system, measures the degree to which the current input matches earlier memory traces. In particular, these earlier traces are the top-down templates (weight vectors) for the F2 (i.e., "field" or layer 2) cells. These F2 cells represent categories and the top-down template corresponds to a description of the prototype of the category. The match computation actually takes place at the F1 cells, which are the input feature representing cells. If the current input is sufficiently close to one of the F2 cell's top-down templates, then that cell becomes active. Thus, the current input is recognized. If the current

input is not sufficiently close to any of the top-down templates, then a new category is established if an F2 cell is available; otherwise, the current input is not recognized.

Although arrived at by different routes, the idea of controlling the embedding of internal representations based on the outcome of a comparison between the system's expected and actual inputs is common to both ART and TEMECOR-II. In general terms, the match process accomplishes the same function in both models-specifically, increased separation of traces in the mismatch condition and increased overlap of traces in the match condition. In the case of the ART models, under the assumption of winner-take-all dynamics at F2 (i.e., singleton category representations), these distinctions are binary. In the mismatch case, the new trace is completely separate from any preexisting trace because a new, previously uncommitted F2 cell has been chosen. In the match case, the new trace is completely overlapped with a preexisting trace, specifically that corresponding to the winning F2 cell.

In contrast, because TEMECOR-II assumes distributed representations at its layer two (L2), there exists a range of possible degrees of overlap between any preexisting internal representation (IR) and the IR being chosen in the current instance. Therefore, rather than having a single threshold for judging the similarity of the current input and the expected input (cf. ART's vigilance parameter), the continuous-valued (between 0 and 1) output of the TEMECOR's comparison process is used to inject a variable amount of noise into the IR-selection process so that continuity between the input layer (L1) and the IR-layer (L2) is achieved. In this sense, TEMECOR-II can be viewed as a generalization of ART to the domain of distributed representations. The issue of continuity is irrelevant to winner-take-all versions of ART since, by definition, IRs have zero overlap.

It is instructive to compare models like ART and TEMECOR-II, which centrally involve the expectation match/mismatch process, to another well-known neural model that does not, backpropagation. Backpropagation utilizes distributed representations and achieves continuity in the input-IR mapping, as evidenced by cluster analyses (Elman, 1990). However, it has limitations with respect to the code stability issue. Specifically, "vanilla" Backpropagation suffers from the previously mentioned catastrophic interference problem in which new memory traces quickly overwrite old memory traces. We note that a number of recent proposals (McRae & Hetherington, 1993; Kortge, 1990; Kruschke, 1992; French, 1991, 1994; McClelland et al. 1994) have been put forth to remedy the problem. Table 1.1 summarizes these models on some relevant features.

Table 1.1: Comparison between ART, TEMECOR and Backpropagation with respect to some issues relevant to code stability.

	ART	TEMECOR	Backpropagation
Competitive Learning	yes	yes	no
Expectation Match/Mismatch	yes	yes	no
Binary Reset	yes	no	-
Graded Reset	no	yes	-
Input-IR continuity	no	yes	yes
Code Stability	yes	yes	no

Chapter 2. Related Work

This chapter is divided into three sections reviewing work related to the current proposal according to three different themes. The first section summarizes other work utilizing the same combinatorial memory scheme that TEMECOR does. The second section summarizes earlier work on the problem of representing temporal sequence information. The final section summarizes some recent combined hippocampal/neocortical models that bear on the issue of episodic vs. semantic memory. We emphasize however that none of the models reviewed simultaneously address the full range of issues that TESMECOR does—i.e., episodic, semantic and temporal sequence memory.

2.1 Other Combinatorial Memory Models

The essential idea of combinatorial memory, which was summarized in Sec. 1.5.1, is eloquently described in terms of the Correlograph model of Willshaw et al. (1969). The purpose of this model is to store associations between spatial pattern pairs, $\{A_i, B_i\}$, such that if A_i is presented by itself at some future time, B_i will be retrieved. The model is depicted in Figure 2.1. Spatial patterns are represented as arrays of pinholes through which light rays can pass. In Figure 2.1a, a diffuse light, D, is turned on and rays shine from the holes in A through the holes in B and land on the plane, C. The pattern of spots on plane C is the *correlogram* of A and B. Pinholes are then drilled at each of the $N_A \times N_B$ spots in the correlogram, where N_A and N_B are the numbers of pinholes in patterns A and B. In order to retrieve A, the apparatus is run backwards. In Figure 2.1b, D is on the left and rays shine from the pinholes in C through the holes in B producing spots at A. The problem is that there are many more possible rays in going from C to B. In fact, there are $N_C \times N_B = N_A \times (N_B)^2$ spots at A. However, only the two spots that were contained in the original pattern A receive three converging rays (solid arrows in 2.1b). The other *spurious* spots (dotted arrows) will, barring coincidences, generally be dimmer than the two correct spots. Pattern A can thus be recovered if we admit only spots which exceed a *recall threshold* set at slightly less than three brightness units (Willshaw et al., 1969).

It should be clear that this threshold-based recall mechanism will allow the same correlogram plane, C, to store multiple pattern pair associations. However, both the number and average intensity of spurious spots (i.e., crosstalk) will increase as we saturate the memory.

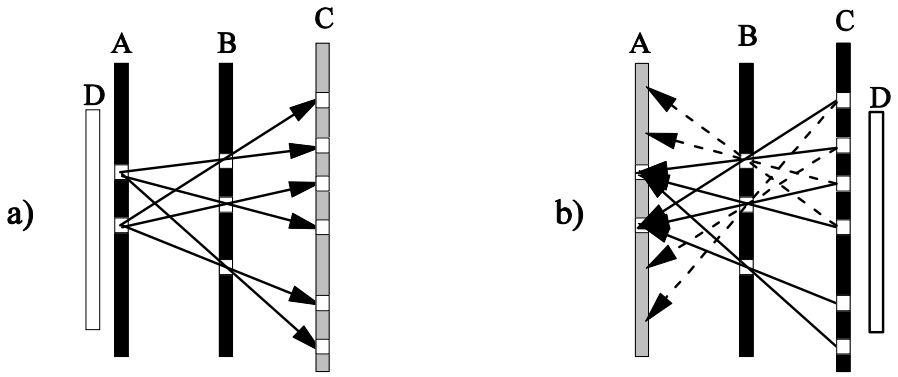

Figure 2.1: The Correlograph model of Willshaw et al. (1969). a) Rays of light pass through holes in A and then through holes in B to produce spots at C. b) Pattern A is retrieved by shining light through C and B onto A. Only the two spots in the original pattern A will have 3 units of brightness. The other spurious spots, resulting from the dotted rays, will be dimmer. A detector with threshold set just less than three brightness units can therefore recover pattern A.

Willshaw et al. (1969) also describe a discrete version of the model above—one that is more plausible from a physiological standpoint. Like TEMECOR, it is an associative network having binary-valued cell activations and binary-valued weights. It also uses a simple *Hebbian* learning rule except that weights are increased when there is *simultaneous*} pre- and post-synaptic activity, rather than *sequential* pre- and post-synaptic activity, as in TEMECOR. The θ parameter of TEMECOR is exactly analogous to the recall threshold used in the Correlograph.

This basic *combinatorial memory* scheme forms the representational basis of many other models as well (Marr, 1969; Lynch, 1986; Miller, 1991; Coultrip & Granger, 1994; Moll, Miikkulainen & Abbey, 1993, Moll & Miikkulainen, 1995). In all of these models, combinations of co-active cells are necessary in order to make another cell fire and these combinations can be highly overlapped as long as the recall threshold is set somewhat higher than the average expected overlap between combinations. Note that all of these models are described only in the spatial pattern domain.

2.2 The Problem of Representing Complex Sequences: Historical Perspective

2.2.1 Lashley and related localist models

The CSS problem is the core of the temporal order problem originally described in (Lashley, 1951). Lashley pointed out that the various constituents that appear in organized sequential behaviors—e.g., the order of phonemes in a spoken word—have no absolute temporal *valence*}. He adduces the example that the same set of phonemes occurs in the words, ‘rite’ and ‘tire’ but in different orders. He makes the same argument in terms of the ordering of words in sentences as well. From examples like this, he concludes, “the order must therefore be imposed upon the motor elements by some organization other than direct associative connections between them” (p.115). That is, Lashley concludes that *associative chaining* is fundamentally unable to handle the CSS problem. Accordingly, Lashley postulates that the determination of order in any particular instance is under the control of the ``idea to be expressed" (p.117) in that instance. Although \Lashley (1951) never describes a precise mechanism whereby the *idea* (or *plan*) imposes the order of read-out of the constituents, he does go on to say that within the nervous system, temporal order must be equivalent, in some sense, to some ‘spatial distribution of memory traces’ (p.128).

A chaining theory is one in which the neural representation that becomes active on time step *t* establishes or strengthens its connection to the neural representation that becomes active on the next time step, *t+1*. Figure 2.2 depicts a concrete example of the problem Lashley identified. The words ‘beet’ and ‘read’ can be phonetically written as [/b/ /i/ /t/] and [/r/ /i/ /d/], respectively (where /i/ is pronounced as a long ‘e’ sound). The /i/ sound must be followed with the /t/ sound when the /b/ sound precedes the /i/ sound, and with /d/ when /r/ precedes /i/. Figure 2.2a shows the initial state of affairs in which neither word, *qua* state sequence, is stored—the synapses are all small. Panel b shows the two synapses that would be increased when the three state (in this case, phonemic) representations, /b/, /i/ and /t/, are activated in sequence. Panel c shows the changes that would occur during experiencing the phonemic sequence, [/r/ /i/ /d/]. Panel d shows the state of affairs after both sequences have been learned. Panels e and f show the essential problem with using associative chaining in conjunction with singleton (i.e., localist)representations. In particular, the two sequences interfere with each other during recall because of the lack of specificity of the *single* cell representing the phoneme, /i/.

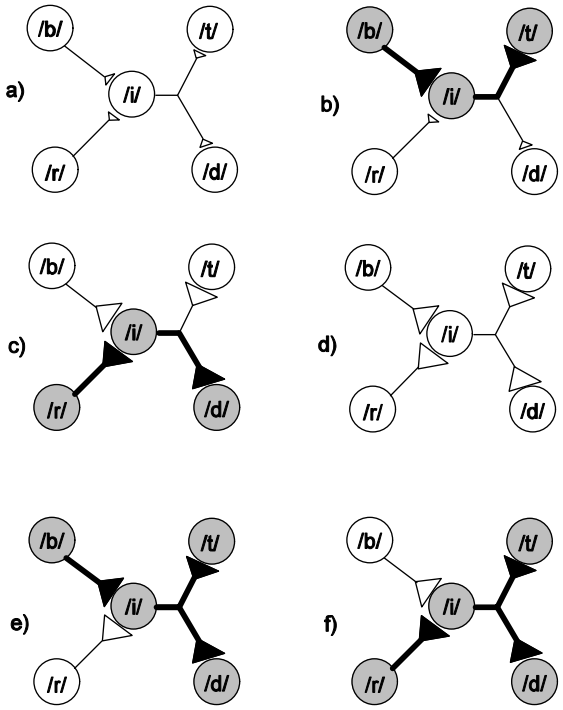

Figure 2.2: A singleton representation for storing the two phonemic sequences for 'beet' and 'read' which have the common middle phoneme, /i/. See text for description of panels a thru f.

One might imagine getting out of the dilemma in this example by assuming a different *single* cell represents the /i/ phoneme in each word. But do we assume a different cell for all instances of the /i/ phoneme? And for every instance of every phoneme? In this case, the model would suffer from *combinatorial explosion*}. In addition, if multiple disjoint representations are assumed for the various instances of /i/, there is the problem of explaining how the *category*}, /i/, is represented; that is, how the invariant aspects of all instances of /i/ are represented.

Lashley speculated that something external to the representations of the constituents of a sequence—which we'll call a *plan* unit—encodes order by some 'spatial distribution of memory traces'. Various more recent neural models addressing issues of temporal order (Grossberg, 1978, 1986; Bradski, Carpenter & Grossberg, 1992; Cohen & Grossberg, 1987) have implemented this general idea in terms of a *gradient* of synaptic weights. The basic idea is depicted in Figure 2.3. Coupled with various assumptions about cell firing thresholds and perhaps auxiliary circuitry for shutting cells off for an extended period once they fire, this 'plan unit' idea can explain how a

particular cell firing sequence can be accomplished when a top-level cell (e.g., the one labeled /beet/) is activated.

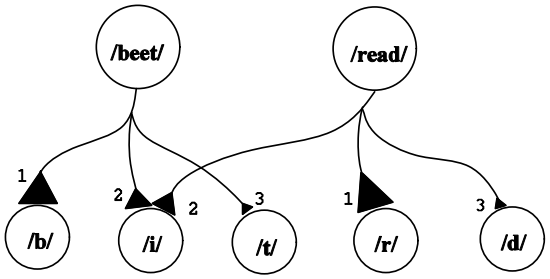

Figure 2.3: This figure shows one possible way of implementing Lashley's plan unit idea. The gradient of LTM weights results in the correct phoneme sequence for each word. Note that additional circuitry (not shown) would be needed to fully implement this idea.

In the development of his theory, Lashley initially hypothesized a mechanism, which included two main ideas: *chaining* and *singleton representation*. In the end, he rejected the chaining component. In Contrast, the TEMECOR theory rejects the singleton representation component, and retains chaining. It will be demonstrated in this thesis how the use of a particular type of *distributed representation*—specifically, a *combinatorial representation*[8]—in conjunction with chaining provides an extremely general and robust solution to the problem of storing, recalling and recognizing highly non-orthogonal spatiotemporal patterns—in particular, CSSs.

In order to place the proposed theory in the context of other theories of serial order, these theories will be characterized on the basis of several properties: (a) whether or not the model uses a singleton or a distributed representation, (b) whether it is a chaining or a non-chaining theory, (c) whether it uses LTM gradients (i.e., *plan* units), and if it is a chaining theory, then (d) whether the cells whose LTM gradients are used are 'in the chain' or 'out of the chain'. In addition, we will be interested in whether or not any particular theory uses the concept of *hierarchical processing* (i.e., *chunking*) or more generally, whether the essential mechanisms of a given theory can be embedded in a hierarchical processing context.

Thus, Lashley's conceptual model would be characterized as a non-chaining model using a singleton representation and LTM gradients.

---

[8] Defined in Ch. 3

Wickelgren (1969b, 1969a) proposed that the central unit of encoding of speech is not the phoneme but rather the *allophone*. He then points out that no two words consist of the same set of allophones. He claims that because the number of allophones is so much larger than the number of phonemes, a chaining theory becomes plausible. However, he still utilizes ‘word representatives’ (i.e., plan units) in his theory and so it is not purely a chaining theory.

Although Wickelgren greatly increased the number of basic units, the more critical fact is that he still assumes a *singleton* representation (of allophones instead of phonemes) rather than a distributed representation. Wickelgren assumes that the theory only needs to account for production of phrases up to perhaps around 100 allophones (which of course are of the same temporal scale as phonemes) in length. From the standpoint of parsimony, the neural chains corresponding to all words in and all potential utterances over a person's lexicon should be embedded in the single set of singleton allophone representations. While it is true that the number of instances of ambiguous branch points in chains (of the kind depicted in Figure 2.2) will be much less under the allophonic (as opposed to the phonemic) representation, such ambiguous points will still occur. The phrases ‘mellow song’ and ‘yellow bird’ for example, have the common allophonic substring, $[_{e}l_{o}\ ,\ _{l}o_{s}]$.

$$\text{mellow song} = {}_{*}m_{e}\ ,\ {}_{m}e_{l}\ ,\ {}_{e}l_{o}\ ,\ {}_{l}o_{s}\ ,\ {}_{o}s_{aw}\ ,\ {}_{s}aw_{n}\ ,\ {}_{aw}n_{g}\ ,\ {}_{n}g_{+}$$

$$\text{yellow seal} = {}_{*}y_{e}\ ,\ {}_{y}e_{l}\ ,\ {}_{e}l_{o}\ ,\ {}_{l}o_{s}\ ,\ {}_{o}s_{E}\ ,\ {}_{s}E_{l}\ ,\ {}_{E}l_{+}$$

In order to handle such sequences, which could clearly be quite frequent when considering the space of all possible phrases of on the order of 100 phonemes, Wickelgren resorted to using plan units to supplement the chains embedded directly in the connections amongst the allophone units.

In terms of the theory classification scheme proposed above, Wickelgren's theory a) uses singleton representations, both for the allophones and for words, b) uses chaining amongst the allophone units, and c) utilizes plan units, which d) are located *external* to the chains.

As mentioned earlier, the operative mechanism underlying Lashley's *plan unit* is a gradient of LTM weights. In Wickelgren's theory, such plan units are located external to his chains but there is no reason why the LTM gradients of cells located *in the chains* cannot also be used to help disambiguate between state sequences. This basic idea is depicted in Figure 2.4 where it is used in the context of a singleton representation. If cell A becomes active on time step t_1 then signals are delivered to synapses s(A,B) and s(A,C) which have weights, w(A,B) and (A,C), respectively.

Since w(A,B) > w(A,C) we can assume that cell B will fire on step t_2. When B fires, it will deliver signals to synapses s(B,C) and s(B,E), which for this example are assumed to be effectively equal. But since cell C also received an input, albeit a smaller one, via s(A,C) on the prior time step, cell C can be assumed to have a higher total input at t_3 than cell E and so will become active. (Note, this explanation requires some additional assumptions—e.g., that cells C and E compete with each other, or that some threshold must be exceeded in order to fire.) A similar argument shows that if we initially activated cell D, then cells B and E would follow.

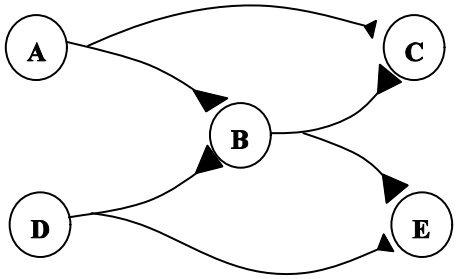

Figure 2.4: This figure depicts a singleton representation in which the LTM gradients of cells located within the chains are used to disambiguate between the two state sequences, [A B C] and [D B E]. See text for explanation.

This type of solution, which is utilized by a number of researchers (Vogh, 1993; Grossberg, 1978), is a means of allowing the prior context of an unfolding state sequence to influence how that sequence unfolds. However it is limited in that it can only implement contextual dependencies that are on the order of the period of time a cell remains suprathreshold when it fires. This period of time within which cells have fired but not yet subsided below threshold and can influence current processing (i.e., the decision as to which cells will fire on the current time step) is referred to as the *window of context*}. The basic idea is depicted in Figure 2.5. Assuming a Hebbian form of learning (in which the amount of weight change is proportional to the product of the activities of the pre- and postsynaptic cells) the average synaptic weight change onto a cell active at time *t* from cells active on progressively earlier time steps falls off quickly.

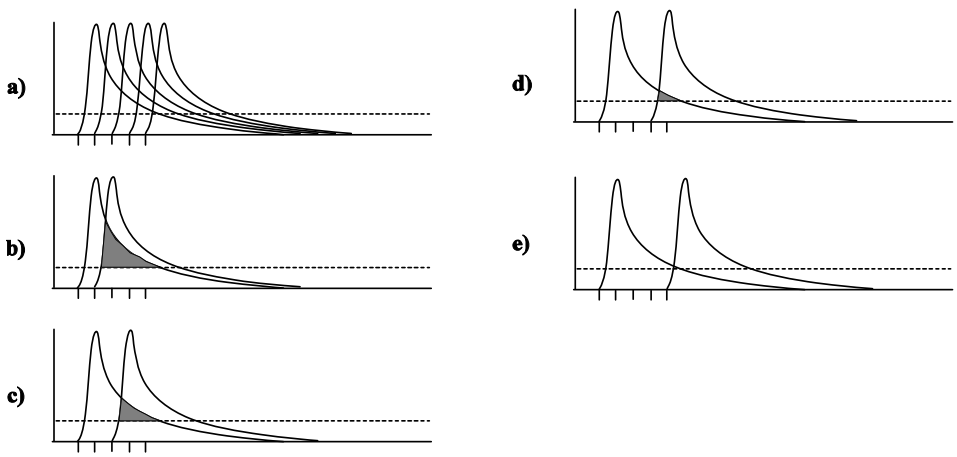

Figure 2.5: This figure illustrates the idea that in order for synaptic growth to occur between the pre- and postsynaptic cell, the supra-threshold portions of their activation profiles must overlap. The tick marks mark off units of time equal to the synaptic delay. Thus in a system having the parameters that would generate these profiles, cells that became active more than three time steps earlier cannot affect current processing decisions.

One possible way to extend this window of context is to have cells remain active for a longer period of time (than that due to the natural activation profile), perhaps by adding recurrent reverberatory loops to the architecture. But this is a nontrivial architectural modification. For instance, how do you decide how long the artificially extended activation period of a cell should last, in general? Furthermore, if such reverberatory mechanisms are added, one then needs to specify how the reverberations are terminated. Even if we assume that such a reverberatory mechanism can be added to the model, there are still problems that seem hard for the 'LTM gradient, in-chain' temporal order mechanism. For example, what if the system must learn the sequences [A B C] and [A C]? Figure 2.6 shows how the weights would be increased on the basis of experiencing these two sequences.

The problem here is that w(A,C) will be increased when sequence [A C] is experienced. Thus w(A,C) will be commensurate with w(B,C). In fact, w(A,C) would probably be greater than w(B,C) because of the increase to w(A,C) that occurs when sequence [A B C] is experienced. The result is that the sequence [A B C] cannot be recalled because cell C will become active at the same time as, or even earlier than, cell B. It is unclear how the addition of reverberatory loops to the architecture can help with this type of problem.

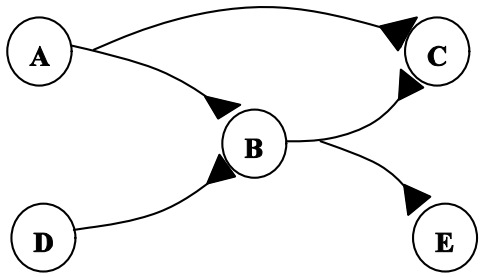

Figure 2.6: The problem here is that the sequence [A B C] cannot be recalled, because cell C would become active at the same time as (or even earlier than) cell B.

Alternatively one could posit the existence of other cells that can represent state A, but this suffers from the two previously described problems: a) *combinatorial explosion*}—i.e., Requiring a unique cell for each instance of the state quickly exhausts capacity, and b) one must explain how the system represents that all those instances are of the same category.

Note that the *Masking Field* of (Cohen & Grossberg, 1987) addresses the type of problem shown in Figure 2.6 but is not a chaining model. The cells in the Masking Field are essentially instances of *plan units*.

Another problem with the 'LTM gradient of cells within the chain' mechanism is that it implies that as we increase the temporal window of context, thus allowing more and more cells to influence any given cell on the current time step, the differences in activation levels upon which the system will make its decision approach the level of neural noise. This is depicted graphically in Figure 2.7.

It is the case in any theory that utilizes LTM gradients for disambiguation (regardless of whether the cells whose LTM gradients are used to disambiguate state sequences are located internal or external to any embedded neural chains that a model may posit) that in order for any cell to influence another cell, it must physically make contact with it. In this sense, the 'window of context' is a structural parameter in such models. Not only must there be a physical connection, but the parameters affecting the cell activation profile must also be set in particular way. In contrast, the window of context is a functional parameter in models in which information about the previous state history leading up to the current moment is implicit in the instantaneous activation pattern—namely, the recurrent BP-based models (Jordan, 1986; Elman, 1990; Williams & Zipser, 1989), the *General Dynamic Model* (GDM) (Ans, Coiton, Gilhodes & Velay, 1994), and TEMECOR.

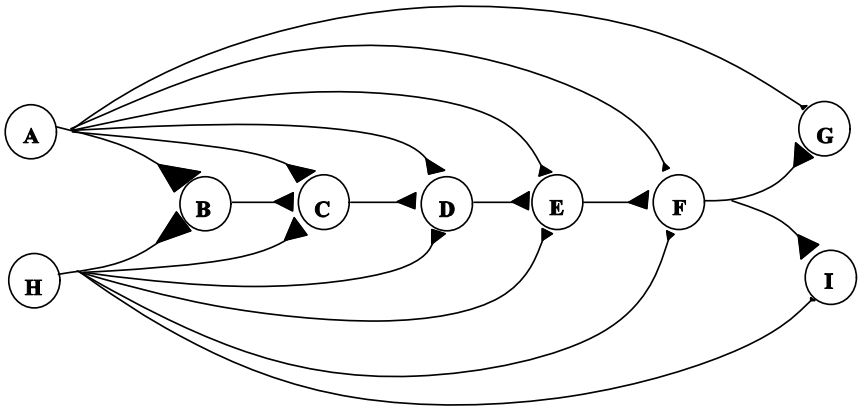

Figure 2.7: Illustration of how the disambiguation signal approaches the level of noise in a singleton representation of sequences. If sequences 1 and 2 are learned, then the only disambiguating influence can be from cells A and H (i.e., the most recent point at which the two sequences were distinct). However, since they were active so many time steps prior to cells G and I, their respective synapses onto cells G and I will be very small under any reasonable Hebbian learning assumption. If recall of seq. 1 is attempted, then by the last step, cell G will have as input, one large input from cell F and one very small input from A. Since cell I's input from F must, by symmetry of the argument, be assumed essentially equal to F's input to G, then G's total input is only slightly larger than I's total input. Note that we left out of the picture (and the argument) the synapses from each of the intermediate cells, B,C,D, and E onto G and I because they provide no differences in input to G and I [just as w(F,G) was assumed essentially equal to w(F,I)] and thus don't change the qualitative argument made in the text that the differences in activation levels upon which the system will make its decision approach the level of neural noise.

2.2.2 Hierarchical processing

Hierarchical processing provides another mechanism for representing sequentiality in neural network models.[9] Figure 2.8 depicts the basic hierarchical processing situation. The idea here is that in order to read out sequence [D,E,F,G,H,I], cell A (which is essentially the singleton representation of that particular sequence) is activated. By virtue of cell A's LTM gradient, cell B will become active next. But note that such a model assumes some sort of inhibitory mechanism (i.e., circuitry) between cells B and C [for example, of the type described in (Grossberg, 1986, p. 222]. Similarly, by virtue of B's gradient, the lowest-level (i.e., state-encoding) cells will become active in the order D, E, F. But again, some sort of competition mechanism at the lowest level must be assumed. More importantly, the parameters (i.e., time scales) of the competitive dynamics of the lowest level of the hierarchy must be different than those of the middle level. Cell B for example, will remain active while cells D, E, and F read out (as a first approximation). Similarly, cell A remains active for the duration of the whole sequence. Thus the relationships between the parameters of the competitive dynamics assumed at each level must satisfy certain relative constraints (Cohen & Grossberg, 1987, the concept of *self-similarity*).

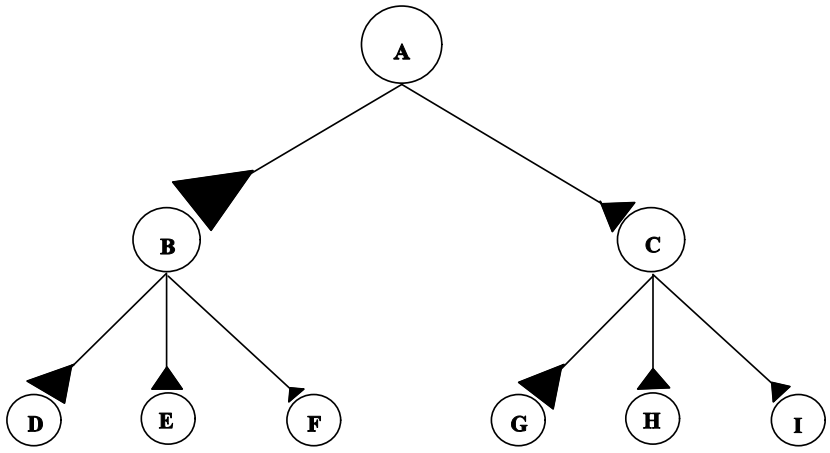

Figure 2.8: Hierarchical encoding of the state sequence [D,E,F,G,H,I]. Note that in this figure, letters A, B, and C do not denote states but are just cell names.

[9] Hierarchical processing again involves the use of the LTM gradient mechanism, but generalized for operation over an arbitrary number of levels.

It was mentioned earlier that one way to extend the window of context in schemes that use an LTM gradient to represent sequentiality is to provide additional mechanisms (i.e., reverberatory loops) for allowing cells to stay on for longer periods of time. However, this non-trivial architectural modification—raises a lot of design questions that are hard to answer with generality. The use of hierarchical processing and in particular, the assumption that cells stay on for longer periods (and generally may have different activation parameters (i.e., self-similarity)) is essentially another way of achieving larger windows of context. TEMECOR does not require this assumption of a range of cell parameters in order to increase the window of context.

2.2.3 Recurrent back-propagation-based models

Reber (1976) showed that when subjects were asked to perform an episodic recall task that involved memorizing strings of letters generated by an artificial grammar [in particular, a finite-state automaton (FSA)], they were later able to recognize novel instances vs. non-instances of grammatical sequences. Reber's findings showed that subjects automatically and unconsciously acquired knowledge of the underlying statistical or similarity structure of the set of stimuli. This type of behavior is referred to as *implicit learning* and, as discussed in Ch. 1, is subsumed by the generalized definition of semantic memory used herein.

Such an artificial *Reber grammar* is depicted in Figure 2.9. Letter strings, for both learning and test trials, are generated by the grammar by entering at the right and following the transitions until state six is reached at which time the exit transition is taken. Decisions at branch points are made probabilistically. Of particular interest is that the grammar produces complex sequences and, more to the point, that the entire set of letter strings comprising the learning trials constitutes a large set of complex sequences.

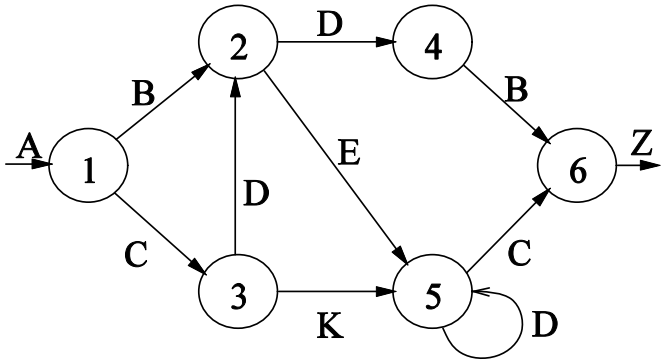

Figure 2.9: A Reber grammar. Processing begins be entering node one. Each arc from any of the nodes has a 50% chance of being traversed. Thus the grammar produces finite state sequences in which states can recur multiple times—i.e., complex state sequences (CSSs).

Recently, recurrent backpropagation-based (RBP) models—in particular, the *Simple Recurrent Network* (SRN) of Elman (1990) and the *Real-Time Recurrent Learning* (RTRL) model of Williams & Zipser (1989)—have demonstrated the ability to embed the statistical structure of the input domain and thus correctly differentiate instances of the grammar from non-instances.[10] This is a spatiotemporal pattern recognition task. The reason the RBP models perform well on this task is because they exhibit the property of continuity in the mapping from the space of inputs (I-space) to the space of internal representations (IR-space). As discussed in Ch. 1, such continuity is necessary if the system is to embed the higher-order statistics of I-space in IR-space. Within the neural network and psychology disciplines, the property of continuity is often referred to as *context-sensitivity*}.

The RBP models achieve this continuity by using context. Specifically, on each time slice, an explicit representation of previous state [either of the output units (JRN) or of the hidden units (SRN)] is combined with the representation of the current state. Figure 2.10 depicts the basic architecture of two of these models. *Hierarchical clustering analysis* reveals that these models have the desirable property (from the standpoint of generalization and categorization) that similarity between internal representations varies directly with similarity of the spatiotemporal contexts in

[10] The recurrent model of Jordan (1986), which we will call the *Jordan Recurrent Network* (JRN), is very similar to the SRN and RTRL and thus would be expected to have essentially the same capabilities and limitations, however it has only been applied to a different type of sequence problem.

which those states occur. For example, all noun tokens cluster in one region of the internal representation space (IR-space), all 'boy' tokens, for example, cluster in a sub-region of the noun region, all *sentence-final* 'boy' tokens cluster in a sub-region of the 'boy' region, etc. Furthermore, using *principal components analysis*, Elman (1991) has shown that such a similarity relationship holds between *trajectories* in I-space and trajectories in IR-space.

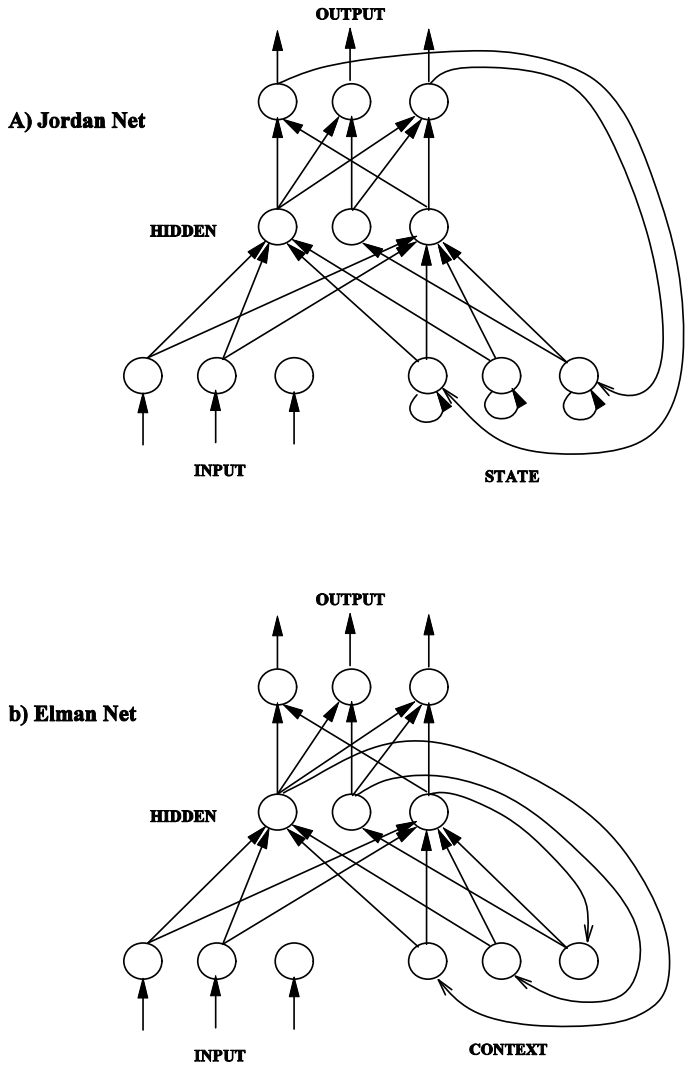

Figure 2.10: Jordan and Elman recurrent nets.

While continuity in the mapping from I-space to IR-space enhances a model's capabilities for similarity-based generalization and categorization, it generally impacts capacity for storing individual items. I am aware of no studies that show such RBP-based models to be capable of

episodically recalling all (or any) of the strings in the training set. Furthermore, these models have great difficulty representing sets of CSSs containing multiple instances of common subsequences (Jordan, 1986; Cleeremans & McClelland, 1990; Smith & Zipser, 1989). These limitations are due to the use of backpropagation, which acts to increase the similarity of [or, 'homogenize'; Cleeremans (1993)] the internal representations corresponding to the various instances of a given state. That is, the same continuity property that leads to these models' success at pattern recognition undermines their abilities at episodic recall. Servan-Schreiber, Cleeremans & McClelland (1991, p. 178) points out that the internal representations (i.e., hidden units patterns) developed by the network reflect two influences: "a 'top-down' pressure to produce the correct output, and a 'bottom-up' pressure from the successive letters in the path which modifies the activation pattern independently of the output to be generated." Assuming that maximal episodic retention corresponds, in general, to minimally overlapped internal representations, it follows that the compression of the space of *actually used* internal representations, due to backpropagation, makes it very unlikely that the RBP models can be adapted so that they exhibit *both* episodic and semantic memory properties.

Furthermore, as these models use backpropagation, they require large numbers of trials per item. In fact, Cleeremans (1993, p.66) states that, ``...the relation between the size of the problem and the number of epochs to reach a learning criterion was exponential for all network sizes."[11]

The basic problem for any of these RBP models—from the standpoint of exhibiting episodic memory for CSSs—is that while it may have a very large internal representation space (IR-space)—i.e., even assuming the cells have only four resolvable levels of activity, a hidden layer of 15 cells has $4^{15} > 1$ billion states (Cleeremans, 1993)—it does not use very much of that IR-space. This has two causes. The first is that backpropagation is not a sparse model—i.e., all nodes and links are involved in representing all input/output pairs. The consequently large degree of overlap between mappings has been cited as a contributing factor in Backpropagation's catastrophic interference problem (McCloskey & Cohen, 1989) and, as McClelland et al. (1994, p. 20) point out, many of the proposed remedies (McRae & Hetherington, 1993; Kortge, 1990; Kruschke, 1992; French, 1991, 1994) "amount to finding ways of reducing overlap of the patterns that are to be

[11] Cleeremans' remarks concern the SRN specifically, however the underlying cause of the problem he identifies is the use of Backpropagation that is common to the JRN and the RTRL models as well.

associated with appropriate responses via connection weight adjustment." The second cause is that the supervised[12] learning procedure, Backpropagation, acts to increase the similarity of [or, 'homogenize'; Cleeremans (1993)] the IRs corresponding to the various instances of a given state.

For example, suppose the JRN is presented with the CSS, [A B C D C E C]. The operation of the JRN is defined so that on each time slice, the target pattern is the next state of the sequence. Thus the same target, C, occurs repeatedly. This imparts a 'force' that pushes the internal representations of the various predecessor states of C (B, D and E) closer and closer together, in terms of Euclidean distance in IR-space, with each additional training trial. The learning algorithm, itself, *tends* to obliterate exactly the temporal context (state history) information that would enable subsequent episodic recall of the sequence. Furthermore, as Cleeremans points out, this effect is only made worse by further learning. This compression of the usable regions of IR-space is due to Backpropagation and therefore occurs in the SRN and the RTRL as well.

Ironically, this 'force' is the same property that accounts for the RBP models' ability to extract the spatiotemporal statistical regularities of the environment (again, as clearly revealed by hierarchical cluster analysis). Therefore there is a fundamental opposition between episodic and semantic memory capabilities in these models.

As can be seen in Figure 2.10, both models use distributed representations, however they are quite different from the combinatorial representation used in TEMECOR. In particular, both the JRN and SRN have a *particular* subset of cells dedicated to representing the last state of the network. Furthermore, neither of these models has an architectural feature analogous to the competitive modules (CMs) of TEMECOR.

Additional problem with RBP models

Although the RBP models can embed general knowledge of the input domain, we must ask how usable is this knowledge? That is, it is not the case that any particular hidden node (or set of hidden nodes) comes to represent any particular feature (i.e., "noun"ness or "boy"ness). In the terminology of Van Gelder (1990), the RBP representational scheme exhibits *functional*

[12] The typical teaching paradigm for these models has been to use state S_{t+1} as the target for S_t. Elman (1991) refers to this as *self-supervised* learning.

compositionality but not *syntactic* (i.e., *structural*}) compositionality. In this case, it seems like it would be hard to add structure (i.e., circuitry) to the model that would bridge the gap from semantic distance encoded only as a completely abstract scalar—i.e., the Euclidean distance in the IR-space—to semantic distance encoded as "lists" or "sets" of features. Yet some sort of bridge like this is needed in order to generate neural network explanations of higher-level cognitive phenomena.

In the RBP models—indeed, in the class of Backpropagation models as a whole—the internal representations for states, and more generally, for state sequences, vary across successive instances (of the states or states sequences) during the learning phase (even though the internal representations may converge asymptotically). This conflicts with evidence regarding the hippocampus (Muller, Kubie, Bostock, Taube & Quirk, 1992; Wilson & McNaughton, 1993) that neural codes remain largely invariant from their point of inception.

2.2.4 Hippocampal model of CSS processing

Levy (1989) provides a detailed discussion of some of the fundamental computational issues involved with representing information, in particular, temporal sequence information, and proposes a sparse distributed model of the hippocampus for processing such information.

Figure 2.11 [from fig.2 of Levy (1989)] graphically summarizes Levy's model. Panel A shows the correspondence between the formal components of the model and the hippocampal circuitry. A central feature of this model is the match computation that is hypothesized to take place at the CA1 field of the hippocampus. The vector of inputs from EC (layer III) to the distal dendrites of the CA1 pyramidal cells represents the current input state. This is matched against the representation of the previous state that has traversed the hippocampal circuit (i.e., EC-DG-CA3) and arrives at the proximal CA1 pyramidal dendrites. TESMECOR also centrally involves a matching process. However it is hypothesized to take place in the CMs that, again, are analogous to minicolumns of entorhinal cortex (EC).

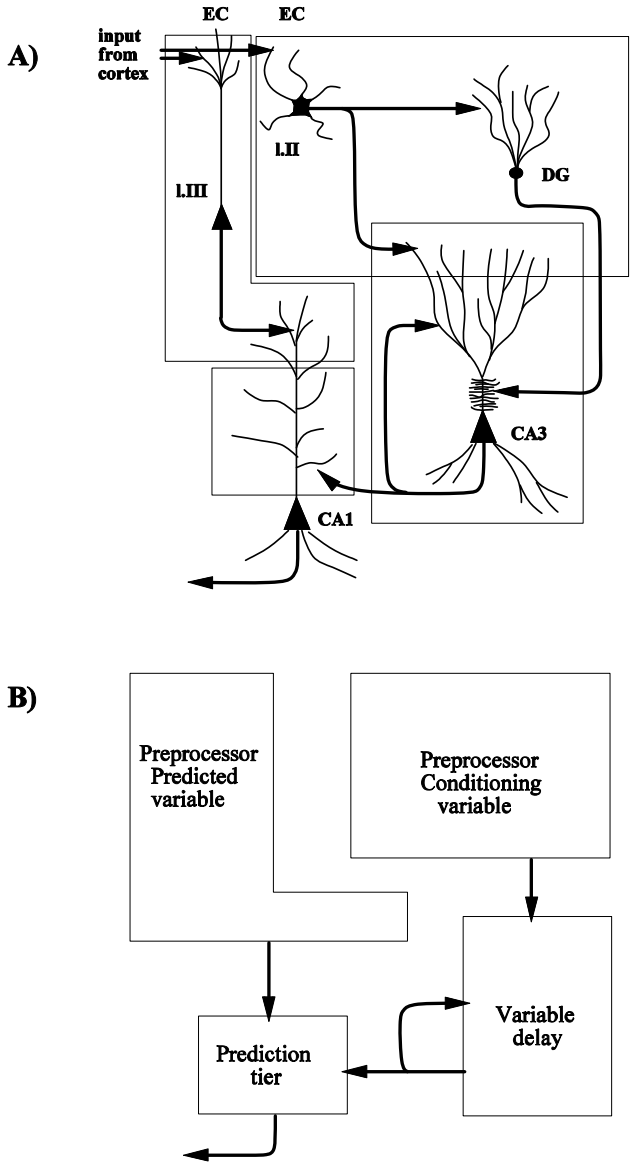

Figure 2.11: A) Summary of Levy's hippocampal model and its correspondence to hippocampal circuitry. B) Summary highlighting just the formal components of the model. This figure is redrawn from fig. 2 of Levy (1989). Copyright ©1989 by William B. Levy.

An important property of his model, as well as of TEMECOR and the RBP models, is that they are strictly local in time. That is, the dynamics of these models depends only on the set of cells active on the previous time slice (i.e., dynamics is first-order in time) and individual nodes do not represent temporal information. All temporal information is such models in present in the macroscopic state—i.e., distributed activation pattern—of the network. In contrast, certain other

models involve either explicit time delays of various magnitudes (Hopfield & Tank, 1989) or capacitive effects (Reiss & Taylor, 1991). That is, the individual elements (i.e., nodes) of the model either must remember information across multiple time slices or must accumulate information across multiple time slices. In either case, individual nodes explicitly store temporal information. Thus, more processing capability is assumed of individual nodes in such models. While these other methods of representing time are probably used within the brain, models involving only first-order temporal dynamics (e.g., Levy's model, RBP models, TEMECOR) reveal an additional, very powerful mechanism for representing sequence information—i.e., that of representing temporal information is the distributed pattern of activity of the network.

As indicated in subsequent reports (Minai & levy, 1993; Minai et al., 1994; Levy et al., 1995; Levy & Wu, 1995; Wu & Levy, 1995), the essential feature accounting for the ability of Levy's model to remember complex sequences is the use of a sparse, distributed representation scheme in conjunction with a learning rule that minimizes overlap between representations. This is essentially what gives TEMECOR its capability to handle CSSs as well. Principled differences in the representational properties of the two architectures cannot be stated in the absence of a comparative study, which is beyond the scope of this report. However, according to several criteria, TEMECOR's demonstrated performance substantially exceeds that of Levy's model. Simulation results for Levy's model have, thus far, only involved small numbers of sequences; often just two, in order to show that two sequences having a common state sequence can be learned. In contrast, the simulations of Sec. 3.6 show thousands of sequences, all chosen from an alphabet of only 100 states being learned. Other simulations in Sec. 3.6 show that very long common subsequences can be learned whereas much shorter common subsequences are attempted by Levy (Minai et al., 1994). Furthermore, most of the reports by Levy and colleagues involve multiple-trial learning of sequences whereas all of the TEMECOR simulations involve single-trial learning. Finally, the input state representations used by Levy and his colleagues have had the constraint that they are all mutually orthogonal, whereas TEMECOR simulations have not been so constrained.

2.2.5 Sliding window models

Elman (1990, p. 180) says the most common approach in the PDP literature to representing time is "the attempt to 'parallelize time' by giving it a spatial representation". In other words, this is the class of models in which separate fields of cells are used to represent different time slices of input,

so that if the model has N such fields, it always explicitly represents the past N time slices of input. The TDNN (Time Delay Neural Network) model of (Waibel, Hanazawa, Hinton, Shakano & Lang, 1989b; Waibel, Sawai & Shikano, 1989) is an instance of this type of approach that I will refer to as *sliding window* solution. Elman lists several problems with the sliding window solution.

a) *It requires an interface to the environment that buffers input.*
b) *The longest (temporally) pattern that can be learned is limited to the size of the buffer, or* shift register*}.*
c) *All input vectors must be the same length.*
d) *It "does not easily distinguish relative temporal position from absolute temporal position."*

2.3 Relation of Combined Cortical/Hippocampal Models

This section summarizes the models presented in McClelland et al., (1994) and O'Reilly & McClelland (1994) and in Murre (1995). These models are similar in that they both concern the relationship of the hippocampal complex to the neocortex. The general thesis, one that is common to other models as well (Rolls, 1989; Alvarez & Squire, 1994), is that the hippocampus stores quickly-learned, but temporary, memory traces which are used to guide the slower embedding of neocortical traces which become permanent. That is, the hippocampus acts as a teacher of neocortex. Both McClelland and his colleagues and Murre point out that this overall scheme provides accounts of both *retrograde amnesia* (RA) and *anterograde amnesia* (AA). Specifically, immediately following initial exposure to an event, the hippocampal trace is strong and the neocortical trace is very weak. Thus, immediate capability for recall and recognition of the event is mediated primarily by the hippocampal trace. With repeated exposures to the event or with internal rehearsal (which may be either conscious or unconscious) of the trace of the initial exposure, the neocortical trace increased in strength. Eventually, the neocortical traces are strong enough so that if the hippocampal complex is subsequently damaged or removed, the organism can still exhibit full recall and recognition of the event. At the same time, the hippocampal trace of the initial event gets weaker and weaker. This may be due to increasing interference from subsequent events or to decay of the traces. The limited size of the hippocampus relative to neocortex suggests interference is a factor and the demonstration of long-term depression (LTD) in both the mossy fibers and the

Schaffer collaterals (Levy & Desmond, 1985; Levy, Colbert & Desmond, 1990) suggests that decay might also be a factor.

Although TEMECOR as presented herein does not specifically model the hippocampus, the addition of a hippocampal analog does solve two important problems for the current version of the model and is a subject of future work. The same general scheme described for the other models will apply to this future version of TEMECOR as well. The reason for summarizing these other models here is that they imply particular theses concerning the relationship of episodic and semantic memory, which differ from that implied by TEMECOR. The most striking difference is that both of these other models are purely spatial models, whereas TEMECOR is a spatiotemporal model.

The second difference is that neither of these theories provides explicit mechanisms whereby expected inputs are matched against actual inputs. Murre (1995) clearly mentions this as a fundamentally important capability of the hippocampal circuitry. However, Murre's model is purely qualitative; no equations or simulations are provided in his paper. Murre's overall scheme would most likely involve his CALM modules (Murre, 1992) that have been defined and analyzed more precisely, however, the overall scheme has not yet been. Although the work contained in McClelland et al. (1994) and O'Reilly & McClelland (1994) is quantitative, there is no explicit mechanism described for matching expected and actual inputs.

In general, the model described in McClelland et al. (1994) seems to be a response to the catastrophic interference problem of the backpropagation model (McCloskey & Cohen, 1989). The failure of traditional backpropagation models to exhibit stable learning of sequentially presented exemplars precludes monolithic backpropagation models from accounting for episodic memory. Essentially, the purpose of the hippocampal subsystem in the model of McClelland et al. (1994) is to turn a sequential presentation schedule in the environment into an *interleaved* presentation schedule for neocortex, which is modeled with a feedforward backpropagation network. In fact, the simulations described in McClelland et al. (1994) use a molar model of the hippocampus. That is, neither the circuitry of the hippocampus nor of its connections to neocortex is actually modeled in this paper.[13] While I agree that one of the main purposes of the hippocampus is to quickly learn new inputs and then to guide the slower embedding of corresponding neocortical traces, the

[13] While the detailed hippocampal model developed in McClelland et al. (1994) apparently does have the desired molar properties assumed in McClelland et al. (1994), a complete simulation modeling all the relevant circuitries would allow a more accurate appraisal of the overall scheme.

particulars of the both the environmental and interleaving schedules used in McClelland et al. (1994) are somewhat at odds with episodic learning. Specifically, each individual input is presented 14 times in a row when it is presented from the environment. Subsequent to the initial presentation from the environment, that input is re-presented from the hippocampus to the neocortex according to a continually declining probability. While the rate of decline seems reasonable, the theory requires traces of all inputs presented up to any point, to remain in the hippocampus and furthermore to remain at full strength in the hippocampus. This suggests that the storage capacity of the hippocampus be at least as big as that of neocortex. One additional potentially problematic assumption of their simulations is that in addition to the set of inputs that are the subject of their experiment, they also pre-train the network with a far larger set of random associations which are also re-presented on each epoch. As is pointed out in French (1991, 1994), pre-training reduces catastrophic interference and it is not clear how much of the performance reported in these simulations is due to the presentations of these other ‘context’ inputs and how much is due to the interleaved training schedule per se.

Chapter 3. The Basic Model: TEMECOR-I

As stated in Sec. 1, TEMECOR-I accounts for some of the more salient, basic properties of episodic memory for the domain of binary, spatiotemporal patterns. In particular, the simulations of Sec. 3.6 demonstrate: very high capacity, single-trial learning, permanence (i.e., stability) of traces, and the ability to store highly overlapped spatiotemporal patterns, including complex state sequences (CSSs). The model can be summarized as an unsupervised, distributed, associative memory whose dynamics are completely local in time and mostly local in space. Additionally, the model satisfies many of the known, fundamental neurobiological constraints. The general neural plausibility of this model is discussed in Sec. 3.5.

As described in Sec. 1.5, TEMECOR-I's great storage capacity stems from: a) its use of a combinatorial representation scheme that provides an exponentially large space of internal representations (IRs); and b) its use of a random method for choosing IRs. This random method of choosing IRs guarantees, statistically, a maximally dispersed—i.e., spread out in the sense of Hamming distance—set of actually used IRs (memory traces). Maximal dispersion over the set of memory traces corresponds to maximal storage capacity.

Table 3.1 provides definitions of the symbols and parameters concerned with TEMECOR-I. All terms are also defined at the point of their introduction in the text. Many of these symbols and terms are carried over to TEMECOR-II. Some symbols have a somewhat different meaning in TEMECOR-II. Table 4.1, in Ch. 4, provides definitions for any additional or re-defined symbols and terms for TEMECOR-II.

3.1 Architecture

TEMECOR-I has two layers as shown in Figure 3.1. Layer 1 (L1)—the input layer—contains M binary feature detectors. Layer 2 (L2)—the internal representation layer—contains M winner-take-all (WTA) *competitive modules* (CMs) which are in one-to-one correspondence with the L1 cells. Each CM has K cells. For simplicity, the mutually inhibitory links among the cells within each CM are not shown. Although the CMs are not explicitly modeled thus far, they could be implemented, for example, in terms of the competitive field theory presented in Grossberg (1973). Whenever a particular L1 cell fires (indicating the presence, in the input, of the corresponding feature), it

enables or activates, via its *feedforward* (F) connections, its corresponding CM. The set of F-connections between L1 and L2 is referred to as the *F-projection*. Exactly one cell in an *active* CM is chosen winner and becomes active. Thus if S L1 cells are active on a given time slice, then S L2 cells will also be active. Both the L1 and L2 representations are distributed, but the L2 representation is much sparser than that of L1. That is, the fraction of cells active in L1 (S/M) is much larger than the fraction active in L2 (S/MK). The L1-to-L2 transform functions similarly to the *mossy fiber-to-granule cell* transform in the cerebellum theories of Marr (1969) and Albus (1971).

As stated in the introduction, each L2 cell has an excitatory modifiable {0,1}-valued synapse onto a proportion, γ, of the L2 cells in each of the other CMs. Except for Sec. 3.6.2, all simulations reported had $\gamma = 1.0$. It is this set of *horizontal connections*—which is called the *H-projection*—in which the chains encoding the temporal aspect of the inputs are embedded. A simple Hebbian learning rule is used (Hebb, 1949). Every L2 cell active at time slice t increases its weight onto *all* L2 cells active at *t+1* unless the weight has already been increased. Thus, it is really more appropriate to think of spatiotemporal *swaths* of activation being embedded in the H-projection rather than *chains*. The use of the term 'swath' here is intended to suggest the width or spatial aspect—i.e., that many cells are active at each successive moment—of the memory trace, as opposed to a 'chain' which suggests a single active unit at each moment.

Each L2 cell has an unmodifiable synapse onto its corresponding L1 cell. As depicted in Figure 1.5 of the introduction, the purpose of these reciprocal (R) or *reverse* connections—which are collectively denoted, the *R-projection*—is to allow the appropriate L1 pattern to be reinstated when an L2 pattern reads out during recall. In TEMECOR-I, neither the F- nor the R-projections are plastic.

Table 3.1: Definitions of Symbols used in TEMECOR.

Symbol	Definition
Γ^p	The layer 1 (L1) spatiotemporal pattern of features for episode, *p*.
Γ_t^p	The spatial pattern of features present on time slice *t* of episode, *p*.
Δ^p	The spatiotemporal pattern of Layer 2 (L2) cell activations for episode, *p*. We may also refer to this as an *L2-code* or as the memory trace for *p*.
Δ_t^p	The spatial pattern of L2 cell activations on time slice *t* of episode, *p*.
M	The number of L1 cells. Also the number of L2 CMs.
K	The number of cells per CM.
L	The Number of L2 cells. $= K \times M$
T	The number of time slices per episode.
S	The number of features present on each time slice of an episode.
STC	A rough measure of the spatiotemporal complexity of an episode. $= S \times M$
E	The number of episodes stored.
Y	The ratio of *E* to *L*.
θ	The number of large co-active synapses that an L2 cell needs in order to become active during recall.
γ	The degree of connectivity in the H-projection.
$\phi_t^p(x)$	The total horizontal synaptic input to cell *x* on time slice *t* of episode *p*. $= \sum_{m \in \Delta_{t-1}^p} w_{mx}$
Λ_t^p	The amount of noise present in the winner selection process on time slice *t* of episode *p*. We can think of it as an input to the cell.
$\rho_t^p(x)$	The total input to an L2 cell, *x*, on time slice, *t*, of episode, *p*. $= \phi_t^p(x) + \Lambda_t^p$

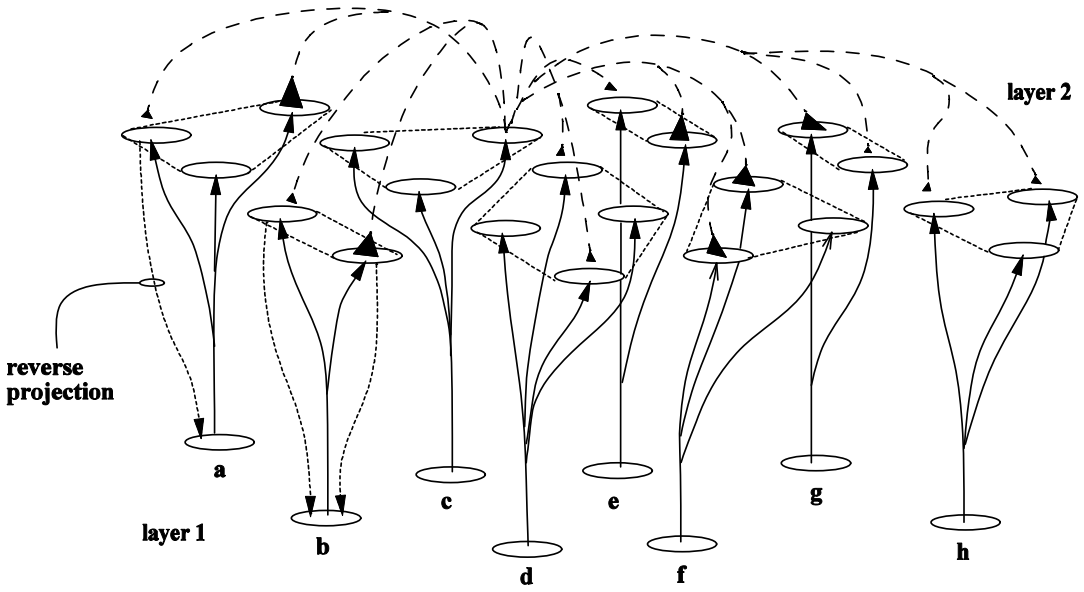

Figure 3.1: TEMECOR-I has two layers. Some of the horizontal connections emanating from one L2 cell are depicted with dashed lines ending in either large (weight = 1) or small (weight = 0) black synapses. Only a few sample reverse (i.e., reciprocal) projections are shown. See text for more explanation.

Figure 3.2 describes the correspondence between two types of views of the model used throughout this work, the plan view and the 3-D view.

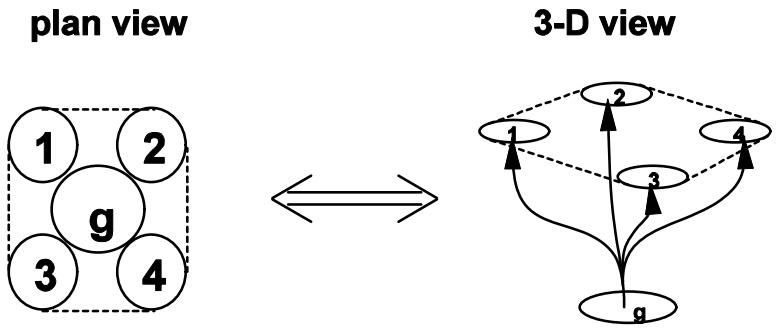

Figure 3.2: Illustration of the correspondence between the two types of views—plan and 3D—used to depict the model, throughout this thesis. The dotted band delimits the cells of a competitive module (CM). The mutually inhibitory links that implement the competition amongst the cells of a CM are not explicitly shown in this or any subsequent figures.

3.2 Notational Format for Episodes (γ-codes) and their Internal Representations Δ-codes

TEMECOR-I requires that environmental states have multiple active features, although it does not require that all states have the same number of active features. However such an assumption does facilitate explanation, so unless otherwise stated, the reader can assume that all states in a given example or simulation have the same number S of active features, where $S < M$. A typical episode, Γ^i, consisting of three time slices can be expressed as:

Γ_1^1:	{a, b, c}		A:	{a, b, c}
Γ_2^1:	{d, e, f}	or,	X:	{d, e, f}
Γ_3^1:	{g, h, i}		B:	{g, h, i}

where each Γ_j^i denotes a particular time slice. Lowercase letters denote features. As shown in the right-hand representation, un-indexed uppercase letters are sometimes used to represent states; this facilitates representing that a particular state occurs in more than one episode and/or more than once in the same episode. The terms episode, spatiotemporal binary feature pattern and state sequence are generally interchangeable in this thesis.

Figure 3.3a shows a possible internal representation, Δ^i, for Γ^i. The terms, 'Internal representation', Δ-code, and *L2 code* are synonymous herein. Δ^i can be written as:

Δ_1^1: $\{a_1, b_2, c_1\}$

Δ_2^1: $\{d_2, e_2, f_3\}$

Δ_3^1: $\{g_4, h_1, i_3\}$

where the notation, a_1, indicates cell 1 in CM_a.

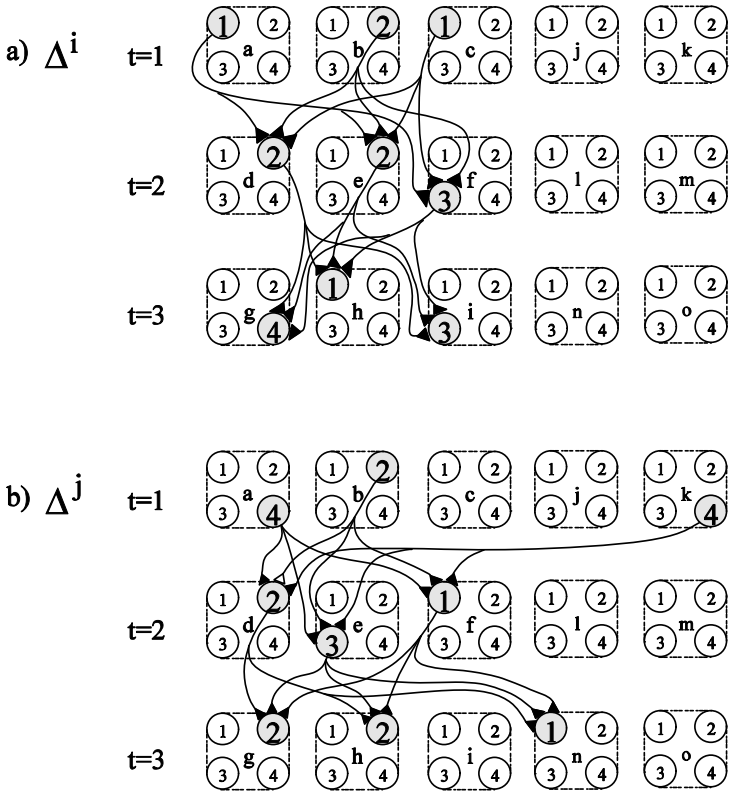

Figure 3.3: Two example L2 codes. a) A particular L2 code (gray cells), Δ^i, that could represent Γ^i. The corresponding learning is also shown. The three rows correspond to three consecutive time steps. b) An L2 code, Δ^j, for Γ^j and the corresponding learning (white (i.e., open) synapses are ones that have been increased in a prior instance; in this case, during Γ^i). Despite a great deal of overlap at the L1 level, the two L2 codes, Δ^i and Δ^j, overlap at only two cells; b_2, on the first time slice and d_2, on the second time slice, and share only one synapse, $s(b_2,d_2)$.

Figure 3.3a also shows the learning that would occur due to presentation of Γ^i. TEMECOR uses the following simple *Hebbian* learning scheme: a synapse, w_{xy} from cell x to cell y is increased to asymptote (i.e., 1) if y is active at $t+1$ and x is active at t. Cell activation levels are {0,1}-valued.

3.3 Processing Algorithms

As mentioned at earlier points in this thesis, TEMECOR-I achieves its great capacity by choosing L2 codes in a random fashion during the learning phase. The choice of winners is not completely random because the set of active CMs is determined by the L1 code (Γ-code). However,

within any active CM, all cells are equally likely to be chosen. This method of choosing winners is referred to as the *random winner choice rule*, (WCR_R).

WCR_R effectively disregards the effects of prior learning in the H-projection. That is, the signals propagating in the H-projection have no influence in determining the winning set of L2 cells. However, they are used on recall trials (as will be demonstrated shortly). Hasselmo, Anderson & Bower (1992) point out that this assumption that the effects of prior learning are suppressed during learning trials—is extremely prevalent in the associative neural network memory literature (Amit, 1988; Anderson, 1972, 1983; Hopfield, 1984; Kohonen 1972, 1988). Hasselmo et al. (1992 refer to this mechanism as *clamping*. In his model, the set of cells comprising the internal representation for each input pattern (Hasselmo's model operates on purely spatial patterns) is selected—i.e., clamped—by the input pattern. In contrast, in TEMECOR-I, the L1 code (i.e., input pattern) chooses only which CMs will be active; the specific winning cell in each active CM is chosen at random.

The assumption that all cells in an active CM are equally likely to be chosen winner in a particular instance is tantamount to the assumption that the amplitude of the noise present during the winner selection process is large relative to the horizontal signal components.

Formally, we can represent the total input, $\rho_t(x)$, to an L2 cell, x, at time t as the sum of two terms: the horizontal synaptic component, $\phi_t(x)$, and noise, $\Lambda_t(x)$:

$$\rho_t(x) = \phi_t(x) + \Lambda_t(x) \qquad (3.1)$$

During learning, $\Lambda_t(x)$ is assumed to be much larger, on average, than $\phi_t(x)$. During recall, $\Lambda_t(x) = 0$ is assumed. Under this assumption, the CMs can be viewed as functioning as WTA modules during learning as well as during recall.

The preceding definition of $\rho_t(x)$ leads to the following learning algorithm for TEMECOR-I.

3.3.1 Learning mode algorithm

On each time slice, *t*, of each episode, *p*, presented during learning:

1. For all L2 cells, x, in *active* CMs, compute $\rho_t(x)$ using Eq. 3.1, where the average value of the noise, $\Lambda_t(x)$, is much higher than $\phi_t(x)$.
2. In each *active* CM, *i*, choose as winner the cell x for which $\rho_t(x)$ is maximal. This set of cells is Δ_t^p.
3. If this is an episode-initial time slice, then do nothing, else increase, to a weight of 1, all horizontal weights from any cell in Δ_{t-1}^p to any cell in Δ_t^p.

3.3.2 Recall mode algorithm

During recall trials, noise is set to zero. The recall threshold, θ, maybe set anywhere from 0 to *S-1*, where the number of L2 cells active on each time slice of an episode is *S*. Maximal capacity is achieved for $\theta = S\text{-}1$. Following the learning phase, if recall is tested for progressively lower values of θ, then progressively more intrusion errors will occur and recall accuracy diminishes. On each time slice, t, of each episode, p, presented during recall:

1. If *t* is an episode-initial time slice, then the precise L2 code corresponding to the first time slice of the episode to be recalled is reinstated. Otherwise, compute: $\phi_t^p = \sum_{m \in \Delta_{t-1}^p} w_{mx}$ for all cells, *x*, in L2. (The ρ values equal the ϕ values since $\Lambda = 0$ during recall.)
2. $\Delta_t^p = \{x \mid \phi_t^p(x)$ is maximal within *x*'s CM, and $\phi_t^p(x) \geq \theta$. If more than one cell within a CM is tied for the maximal ϕ value, then one of them is picked at random.

3.4 Example of Operation

Now suppose another episode, Γ^j, defined as:

Γ_1^1:	{a, b, k}		C:	{a, b, k}
Γ_2^1:	{d, e, f}	or,	X:	{d, e, f}
Γ_3^1:	{g, h, n}		D:	{g, h, n}

is presented. Γ^j is non-orthogonal to Γ^i. In fact, they have exactly the same middle time slice—i.e., $\Gamma_2^i = \Gamma_2^j$. Neither state sequence by itself is complex, but taken together they constitute a set of CSSs. The use of WCR_R ensures, statistically, that the overlap between any two L2 codes—even ones corresponding to the same set of L1 features (i.e., state)—will be small. The probability that the two L2 codes, Δ_2^i (= Δ_X^i) and Δ_2^j (= Δ_X^j), will be identical is:

$$p(\Delta_X^i = \Delta_X^j) = K^{-s} \tag{3.1}$$

where *S* is the number of active features in X, and *K* is the number of cells per CM.

More generally, let *Z* be the number of cells in common between two L2 codes for a given state, X. That is, $Z = \left|\Delta_X^i \cap \Delta_X^j\right|$. Then, in general, the probability that $Z = q$, for $0 \le q \ge S$ is:

$$p(X = q) = \binom{S}{q}\left(\frac{1}{K}\right)^q\left(\frac{K-1}{K}\right)^{S-q} \tag{3.2}$$

This is the binomial distribution with $p = 1/K$ and $q = 1 - p = (K-1)/K$ and its expected value—i.e., the expected overlap, $\hat{Z}$—is just *S/K*.

Accordingly, the L2 code, Δ^j, which appears in Figure 3.3b, has very little overlap with Δ^i. The L2-overlap, $\Delta^i \cap \Delta^j$, has only two cells, b_2 and d_2, whereas L1-overlap, $\Gamma^i \cap \Gamma^j$, has 7 cells. Thus the L1-to-L2 transformation serves to *separate* the patterns. This basic property of pattern

separation has been used for the purpose of maximizing capacity in many other neural network models including ones concerning the cerebellum (Albus, 1971; Marr, 1969), the hippocampal complex (Shapiro & Olton, 1994; McClelland et al., 1994; O'Reilly & McClelland, 1994), the olfactory/hippocampal system (Lynch & Granger, 1994), and general neocortex (Moll et al., 1993; Moll & Miikkulainen, 1995).

The low expected overlap, $\hat{Z}$, suggests that memory traces can be kept from interfering with each other *during recall* by requiring that a cell have at least θ active, large (i.e., a weight of 1) synapses in order to fire. The parameter, θ, is called the *recall threshold*.

Figure 3.4 shows that Γ^i is recalled perfectly if $S \geq \theta \geq 2$. Cells e_3 and f_1 receive input only from b_2 and so do not meet θ and remain inactive. Cells, d_2, e_2 and f_3, receive three large inputs and correctly become active on the second time slice of this recall trial. Similarly, none of g_2, h_2 and n_1, become active at $t = 3$ because none of them meets θ.

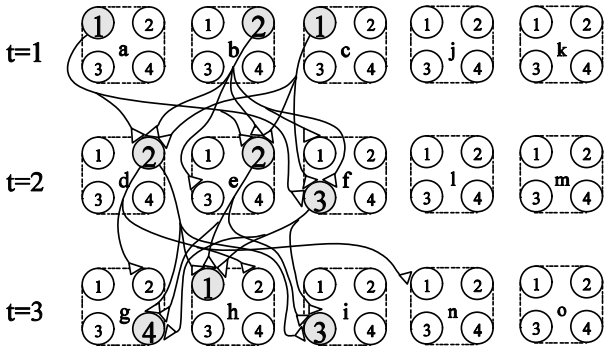

Figure 3.4: Recall of Γ^i in the case of $3 \geq \theta \geq 2$. If $\theta = 1$, then n_1 would become active at t = 3. If $\theta > 3$, then no recall at all is possible.

Figure 3.5 shows more clearly the situations that exist at $t = 2$ and $t = 3$ during attempted recall of Γ^i. The spurious inputs to the inappropriate cells are shown and it can clearly be seen that the cells that should not become active do not exceed θ for $3 \geq \theta \geq 2$.

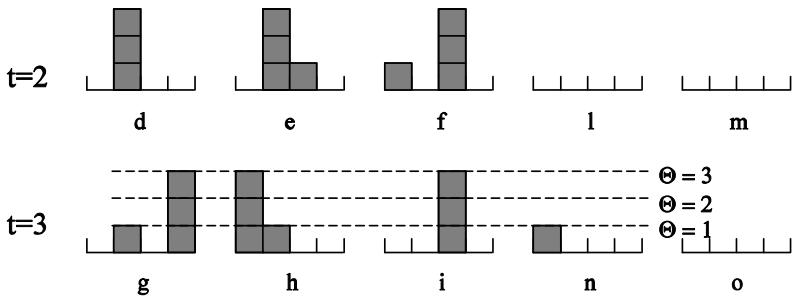

Figure 3.5: Another depiction of the inputs (to all L2 cells) that exist at t = 2 and t = 3 when the model has been prompted with the first time slice Γ^i **.**

Figure 3.5 suggests a simple explanation for the great capacity of TEMECOR-I. In terms of the figure, it is simply that as more and more episodes are experienced and stored, the heights of the bars representing the total horizontal input to the incorrect cells, for any particular time slice during an attempted recall, increase very uniformly. Figure 3.6 illustrates this in more detail. Figure 3.6a shows a hypothetical histogram of total horizontal inputs to the L2 cells of six CMs (each having 10 cells) that might exist on some recall time slice that obtains after the model has experienced very few episodes—that is, early in the 'life' of the model. This is analogous to the situation depicted in Figure 3.5. Note that in this case, a large range of θ values would yield perfect recall.

Figure 3.6b shows the same time slice during recall of the same hypothetical episode after many more episodes have been learned. There is still a fairly large range of perfect-recall-yielding θ values. Note also that although the heights of the bars corresponding to the incorrect cells are higher, the variance of these heights is rather small compared to the height of the bars corresponding to the winning (i.e., correct) cells. Finally, Figure 3.6c shows the same hypothetical recall situation much later in the life of the model, after many more episodes have been experienced. Perfect recall is still possible but only over a much-reduced range of θ. The capacity of the model derives from the fact that the variance of the total horizontal inputs to incorrect cells remains rather small compared to the maximal horizontal input, across the entire life of the model; in particular, even as it approached saturation (i.e., panel c of the figure). The range of θ values yielding perfect recall are between the dotted lines in panels, a and b, and equal to the dotted line in panel c.

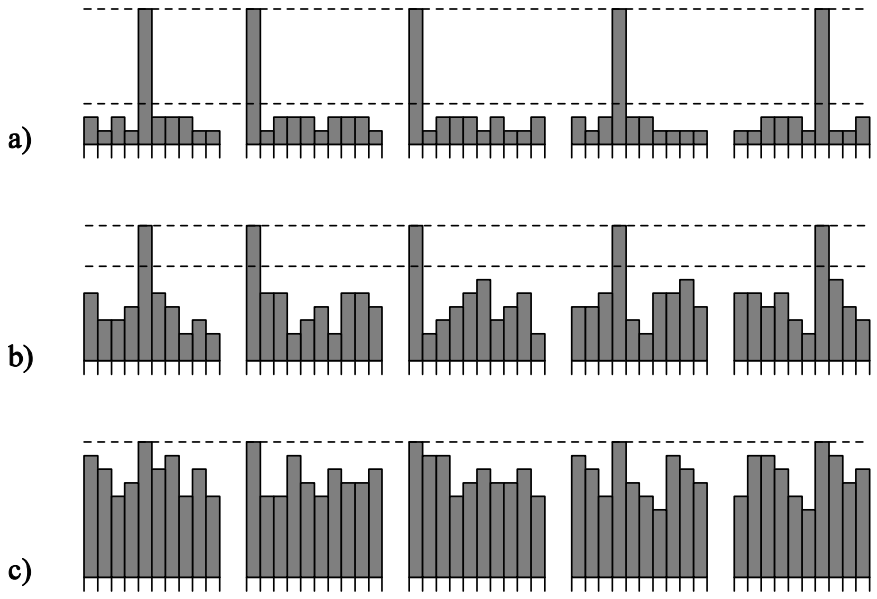

Figure 3.6: Graphical explanation of underlying principle explaining the model's high capacity. a) Histogram of total horizontal inputs that exist during some time slice of a hypothetical recall event early in the 'life' of the model—i.e., after very few episodes have been experienced. b) Same hypothetical recall situation, but at a later period of 'life', after many more episodes have been experienced. c) Again, the same recall event but at a point in the 'life' of the model in which it has neared saturation. Notice that the range of θ values yielding perfect recall shrinks towards its upper limit as saturation increases. See text.

3.5 Possible Neural Interpretation

As stated in the introduction, while functionality has been the primary goal for TEMECOR, it has been inspired by neuroanatomy and neurophysiology. In those cases where a particular feature of the model does deviate from what is known of the actual brain, the deviation is generally of a quantitative nature. For instance, most simulations reported in Sec. 3.6 involve full (i.e., 100%) horizontal connectivity over the entire L2. It is known that the actual degree of horizontal connectivity, over cortical regions encompassing more than several hundred minicolumns, is much lower than that, and that there is distance-dependent fall-off (Szentagothai, 1975). While the qualitative faster-than-linear capacity scaling is preserved in simulations involving horizontal connectivities as low as 70%, the rate of increase falls off rather quickly and may imply actual capacities that are quite small at neurobiologically realistic horizontal connectivities. Two potential

remedies to this problem with the theory are mentioned below. The first is elaborated in the context of a speculative extension to the model of a hippocampal analog in Sec. 4.13. The second remedy has not been developed yet and is a subject for future research.

a) The model's horizontal links may correspond to multi-synaptic paths in the actual brain. In this case the proportion of cells reachable in di- or tri-synaptic paths is much larger than that reachable in a single synapse. This is a non-trivial change to the theory and has not been considered in any detail yet.
b) The basic horizontal linking of high-dimensional features into spatiotemporal traces may be true to the actual neurobiology, however this may take place only over relatively small regions (i.e., patches) of neocortex. In this connection, a key area for future research is to model distance-dependent fall-off in the horizontal connection matrix.

In addition, the model may contain features which do not correspond to actual neurobiology but which are also not crucial to the model. For example, the L1 cells and the L2 CMs are in 1-to-1 correspondence in TEMECOR-I. This constitutes a very strong wiring constraint that is not supported by the known anatomy of the brain. However, as will be seen in Ch. 4, this constraint is greatly relaxed in TEMECOR-II. Thus, this constraint was a theoretical starting point that enabled a certain amount of development but which was then realized to be a non-essential element of the general theory.

From its inception, the TEMECOR model has been envisioned as being analogous to deep neocortex. More specifically, the two layers, L1 and L2, are considered to be analogs of two adjacent cortices with L1 projecting to L2. Thus, L1 might correspond to area TE in inferotemporal cortex, in which case L2 would correspond to entorhinal cortex. The competitive modules (CMs) of L2 are intended to be analogous to the *minicolumns* of neocortex (described shortly). While neither TEMECOR-I nor TEMECOR-II have CMs in L1, this organizational aspect could be added without requiring any qualitative changes in the theory.

The basic operational and organizational principles of the model are general enough so that, with various changes in parameter settings (e.g., relative proportions of horizontal and vertical connectivity, learning rates, etc.) it may be applicable to all of neocortex including primary sensory cortex, in which case L1 would correspond to a thalamic nucleus and L2 to the primary sensory

cortex. However, to reiterate, thus far, TEMECOR has primarily been conceived as analogous to deep cortex.

The minicolumn (Eccles, 1981; Szentagothai, 1975; Mountcastle, 1978) is a roughly cylindrical group of cells oriented perpendicular to, and extending across all, layers of the neocortical sheet. Shaw, Harth & Scheibel (1982, p.337) estimate that a minicolumn contains about 30 pyramidal cells. When Rockel, Hiorns & Powell (1980) counted all cell types, a remarkably consistent (across widespread regions of cortex) figure of 110 ± 10 cells per minicolumn resulted. The L2 cells of the model are proposed as analogs of the layer 2 and layer 3 (i.e., *supragranular*) pyramidals of the cortical minicolumn. This leaves roughly 80 cells per minicolumn to help implement the competition (which is not explicitly modeled within the theory so far) and various support functions like implementing/modulating the recall threshold (θ), changing overall operational modes, etc.

Hubel & Wiesel (1968) found that all of the principal cells within a given minicolumn in V1 (i.e., early visual cortex) have approximately the same receptive field properties, a central feature of the basic model, TEMECOR-I. Gross, Rocha-Miranda & Bender (1972) have found that the IT and prefrontal cells respond to highly complex features. Various researchers have even found that certain cells in IT appear to respond in a highly specific manner to pictures of faces (Bruce, Desimone & Gross, 1981; Perrett, Rolls & Caan, 1982). More recently, Fujita, Tanaka, Ito & Change (1992) reported receptive field homogeneity within the mini-columns of the very deep, anterior IT (inferotemporal) cortex.[14]

The idea that representational cells—in particular, the L2 cells of the proposed theory—respond to specific *conjunctions* of input features is extremely widespread in the memory modeling literature [cf. the *conjunctive encoding* of O'Reilly & McClelland (1994), the *configural memory* of Sutherland & Rudy (1989), and the *relational representations* of Shapiro & Olton (1994)]. This specificity is achieved in TEMECOR by the use of the recall threshold. According to the proposed theory, a given L2 cell will still fire correctly even if a certain proportion, determined by the current setting of the recall threshold, of the L2 cells that should have been active on the prior time slice (i.e., that are contained in the prior L2 context) fail to become active. O'keefe & Conway (1978)

---

[14] Note that due to TEMECOR-II's relaxation of the 1-to-1 correspondence between L1 and L2, the receptive fields of different L2 cells within the CM need not be the same.

and O'keefe & Nadel (1978) have shown that the *place cells* of CA1 respond to conjunctions of input features and that one or several features can generally be deleted without substantially affecting the firing pattern of the cells.

From the standpoint of efficiency, it makes the most sense to form memory traces of complex events involving many high-dimensional features (e.g., a particular person's face) by forming connections amongst units that represent those high-dimensional features. That is, the same information that is present in the spatiotemporal pattern of activity over a field of high-level feature detectors may also be present in the spatiotemporal activity pattern over one or more fields of lower-level feature detectors that feed into the high-level field. However, far less (in principle, exponentially less) physical connections are necessary for representing a concept at the higher-level because there are many less (e.g., polynomially less) representational units involved than in the lower-level cortices. Thus it is much more efficient to encode memories/concepts—*qua* spatiotemporal feature patterns—at the highest level possible. There is much evidence that, in general, the complexity (i.e., dimensionality) of receptive fields of cells increases from earlier to later (i.e., deeper) cortex (Hubel & Wiesel, 1968; Gross et al., 1972; Fujita et al., 1992; Kobatake & Tanaka, 1994; Tanaka, 1993; Hasselmo, Rolls & Baylis, 1989).

Another very important point concerning the neurobiological interpretation of the theory is that the enhanced version, TEMECOR-II, constitutes a specific and detailed theory of how the spatial and temporal aspects of a cortical cell's receptive field are combined to define a single spatiotemporal receptive field. This is an area in which there has been relatively little work. Most of the work in this area has concerned purely spatial receptive field properties.

A central feature of the model is the Hebbian learning from cells active on one time step onto cells active on the next. This is also a central feature of numerous other neural models and has received much experimental support (Levy, 1985; Wigstrom, Gustaffson, Huang & Abraham, 1986; Kelso, Ganong & Brown, 1986).

Although not discussed in detail in this thesis, the *reverse projections* are a fundamental feature of the model. Without them, there is no way to cause an L2 memory trace to elicit the contemporaneous read-out of the corresponding L1 trace. There is ample evidence for the existence of reverse projections (Rockland & Pandya, 1979; Maunsell & Van Essen, 1983) from virtually any region of cortex back to those regions that project to it.

3.6 Simulation Results

Table 3.2 gives the maximal capacity (as well as other statistics) for networks of increasing size, in the case of uncorrelated patterns. All episodes presented in the simulations reported in Table 3.2 had $T = 10$ time slices and each time slice had $S = 20$ (out of $M = 100$) active features, chosen at random. The product, $T \times S$, which is the total number of *featural instances*[15] comprising an episode, will be referred to as the *spatiotemporal complexity* (STC) of an episode. All episodes associated with Table 3.2 have STC = 200. In addition, the number of L1 cells (i.e., features), M, equals 100 for all simulations of Table 3.2.

Furthermore, the recall threshold, θ, is set to $S\text{-}1 = 19$ for all these simulations. The degree of overlap between the L2 codes increases as additional episodes are presented—i.e., as the memory is *saturated*. Thus, maximal capacity is achieved by setting θ as high as possible (but of course, less than S or else no recall is possible).[16]

Table 3.2 was generated in the following way. For each CM size, K, the maximal number, E, of episodes that could be stored to criterion accuracy of approximately 97% was determined.[17] Recall accuracy, $R(e)$, for a given episode e, is defined as:

$$R(e) = \frac{C(e) - D(e)}{C(e) + I(e)} \tag{3.3}$$

where $C(e)$ is the number of L2 cells that should become active during recall of e, $D(e)$ is the number of L2 cells which should have become active but did not (*deletions*)[18], and $I(e)$ is the number of L2 cells which should not have become active but did (*intrusions*). Recall accuracy for a

---

[15] The term, *featural instances* is used here because features can occur more than once in an episode.

[16] Generally, greater recall accuracy is achieved for $\theta = S\text{-}1$ than for $\theta = S$ because any given feature f can occur on consecutive time slices. Recalling that there are no intra-CM excitatory connections, it follows that during the learning trial, the L2 winner for the second instance of f, will have only $S\text{-}1$ (not S) of its incoming synapses increased. Since this event is expected to occur rather frequently, especially as the ratio of S to M increases, maximal capacity is achieved for $\theta = S\text{-}1$.

[17] Each line (i.e., data point) of all tables represents the average of three simulations with the corresponding parameter set.

[18] The term ‘omissions’ is more typical than ‘deletions’ in the experimental psychology literature.

whole set of episodes, R_{set}, is just the average of the R values. All episodes were presented only once.

Table 3.2: Results of simulations using uncorrelated patterns. See text for discussion. All Simulations had θ = 19, S = 20 and T = 10. Abbreviations: E = maximal number of episodes which could be stored to criterion accuracy of approximately 97%; Υ ***is the ratio of episodes E to number of L2 cells L;*** $\hat{F}$ ***= average number of instances of each feature, across entire set of episodes; K = CM size; L = total number of L2 cells;*** $\hat{V}$ ***= average number of times each L2 cell is used; var(V) = variance of V;*** W_{inc} ***= total number of increased weights;*** R_{set} ***= recall accuracy over the whole set of episodes; and H = percentage of horizontal weights increased.***

E	Υ	*F*	*K*	*L*	$\hat{V}$	*Var(V)*	W_{inc}	R_{set}*(%)*	*H(%)*
129.3	0.162	258.7	8	800	32.33	42.78	327465.7	97.8	51.7
290.3	0.242	580.7	12	1200	48.39	58.61	736201.3	96.6	51.6
517.0	0.323	1034.0	16	1600	64.42	78.58	1309367.0	97.2	51.7
793.0	0.397	1586.0	20	2000	79.3	72.69	2020240.0	97.2	51.0
1141.7	0.476	2283.3	24	2400	95.14	86.94	2908397.0	97.1	51.0
1544.7	0.552	3089.3	28	2800	110.33	98.93	3942919.0	97.2	50.8
2002.3	0.626	4004.7	32	3200	125.15	115.21	5122921.3	97.1	50.5
2506.0	0.696	5012.0	36	3600	139.22	177.57	6433038.0	97.1	50.1
3084.0	0.771	6168.0	40	4000	154.2	193.86	7925805.0	97.7	50.0

Table 3.2 supports the claim, made in the introduction, that the number of episodes that can be stored to criterion recall accuracy increases faster-than-linearly, at least over the range of network sizes analyzed, in network size, L.[19] This can also be seen in the curves of Figure 3.7. The four dotted curves correspond to series of experiments involving uncorrelated episodes. The lowest dotted curve of the figure is derived from Table 3.2. The other three (progressively higher) dotted curves of Figure 3.7, correspond to simulations involving episodes with progressively lower STCs—that is, to episodes containing less and less information: STC = 160 (*T = 8, S = 20*), STC = 120 (*T = 6, S = 20*), and STC = 80 (*T = 4, S = 20*).

[19] To be more precise, the size of the network is really the total number of L2 cells, *L*, *plus* the total number of L1 cells, *M*. However, as larger and larger CMs are used, the number of L1 cells becomes insignificant compared to the number of L2 cells.

The solid curves of Figure 3.7 (the lowest one of which is derived from Table 3.3) show that a slightly slower, although still faster-than-linear, relationship also holds for the case of correlated patterns. The episodes—i.e., complex state sequences—used in the simulations of Table 3.3 were constructed as follows. First, a set (alphabet) of $U = 100$ unique states, each consisting of 20 active features, was built. The $T = 10$ time slices composing each episode were then randomly chosen (with replacement) from this alphabet of 100 states.

Figure 3.8 shows, for the same eight simulations of Figure 3.7, a linear relationship between the size of the CMs, K, and the number of times, V, an L2 cell can be used to represent an instance of its corresponding feature *while still meeting the required recall accuracy*. Thus as CM size (and thus total network size) increases, the capacity of individual cells also increases. A systematic study of the relationship of *V* to *K*, while total network size remains constant has not yet been done.

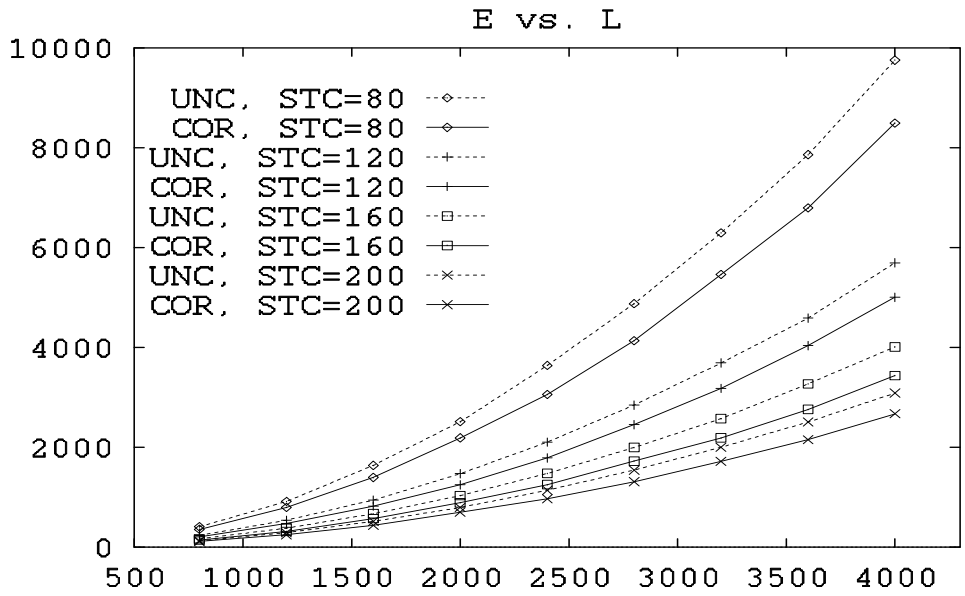

Figure 3.7: The relationship of the number of correctly recalled episodes, E (ordinate), to the number of L2 cells, L (abscissa), for various types and sizes of episodes. The four dotted curves correspond to simulations in which uncorrelated episodes of varying STCs—80, 120, 160 and 200—were used. Solid curves correspond to simulations involving correlated episodes. This figure shows: a) capacity increases faster-than-linearly with the number of cells for both uncorrelated and correlated episodes, b) capacity is higher for uncorrelated episodes than for correlated (for a given STC) and c) capacity increases more quickly as we consider smaller episodes.

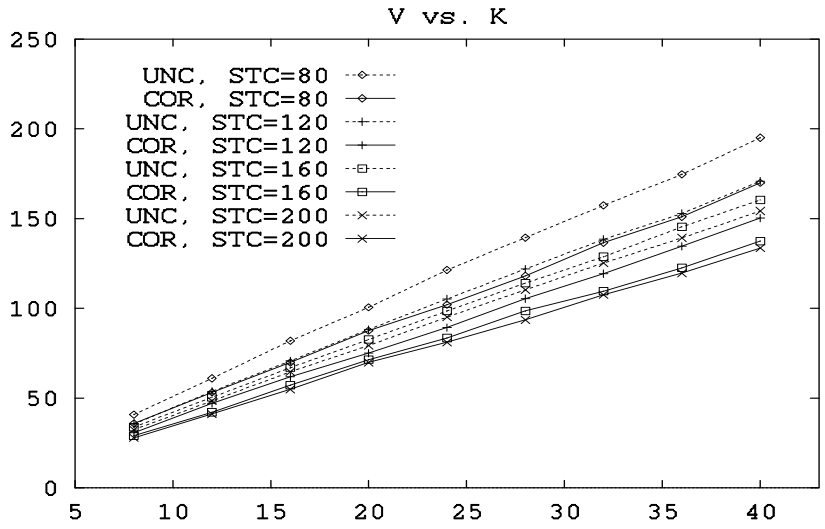

Figure 3.8: The number of times V, an L2 cell can be used to represent an instance of its corresponding feature while still meeting the required recall accuracy.

Table 3.3: Capacity results for the correlated patterns (CSS) case. Abbreviations: E = maximal number of episodes which could be stored to criterion accuracy of approximately 97%; Υ ***is the ratio of episodes E to number of L2 cells L;*** $\hat{F}$ ***= average number of instances of each feature, across entire set of episodes; K = CM size; L = total number of L2 cells;*** $\hat{V}$ ***= average number of times each L2 cell is used; H = percent of horizontal weights increased;*** R_{set} ***= recall accuracy over the whole set of episodes;*** $\hat{Q}$ ***is the average instances of each state, across entire set of episodes.***

E	Υ	*F*	*K*	*L*	$\hat{V}$	R_{set}*(%)*	*H(%)*	$\hat{Q}$
111.7	0.14	223.3	8	800	27.92	96.4	45.8	11.2
246.7	0.206	493.3	12	1200	41.11	97.4	45.1	24.1
439.3	0.275	878.7	16	1600	54.92	97.0	45.2	43.8
698.3	0.349	1396.7	20	2000	69.83	97.3	45.9	70.4
971.7	0.405	1943.3	24	2400	80.97	97.3	44.7	96.8
1309.3	0.468	2618.7	28	2800	93.52	97.2	44.3	130.9
1719.7	0.537	3439.3	32	3200	107.48	97.0	44.6	177.3
2151.7	0.598	4303.3	36	3600	119.54	96.9	44.2	212.7
2671.3	0.668	5342.7	40	4000	133.57	97.0	44.5	286.6

Figure 3.9 plots the ratio, Y = *E/L*, of episodes per L2 cell, stored to criterion recall accuracy of approximately 97%, against the number of L2 cells, *L*. Again, an apparently linear relationship holds. The slope of the lowest dotted graph is about 0.00019 which means that for a network having 100,000 cells, approximately 1.9 million episodes of STC = 200 can be stored.

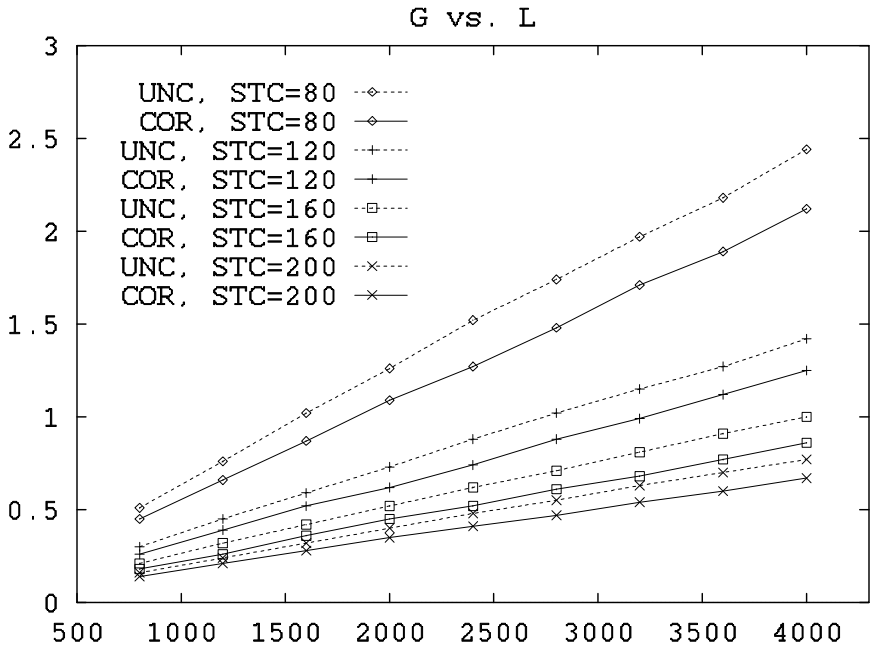

Figure 3.9: Plotting the number Y ***of episodes per L2 cell against the number of L2 cells, L, yields approximately linear relationships in both the correlated and uncorrelated cases.***

3.6.1 Variation of S parameter

The simulations reported in this section are designed to show the specific effect of the *S* parameter—that is, of the number of active features per time slice. All simulations used uncorrelated episodes with STC = 600, with the number of time slices per episode *T* varying inversely with *S*. The parameters corresponding to the six curves plotted in Figure 3.10, in descending order, are:

S	***T***	***H(%)***
20	30	50.0
25	24	57.0
30	20	62.5
40	15	70.0
50	12	74.7
60	10	77.7

where *H* is the percentage of increased horizontal weights.

TEMECOR's parameter *S* corresponds to a parameter known as the *coding rate*. The coding rate is the fraction of the cells of a network that becomes active in a given pattern (Nadal & Toulouse, 1990). TEMECOR's L2 contains $M \times K$ cells. Thus, the coding rate is S / KM and varies across a relatively small range from 20/4000 (0.5%) to 60/4000 (1.5%) for the six curves of Figure 3.10. The rather small change in capacity seen across the six curves in the figure reflects the relatively small variation in coding rate.[20] Presumably, the curves would continue to increase for smaller and smaller *S*. For example, for the sparse spatial associative memory examined in (Palm, 1980), maximal capacity was achieved when patterns consisted of only two or three active cells.

[20] Note that the top curve of Figure 3.10 (for *S = 20, T = 30*) only reaches a height of about one third that of the STC = 200 curve in Figure 3.7 because about three times more information is being stored per episode in this case (i.e., STC = 600) than in the earlier figure.

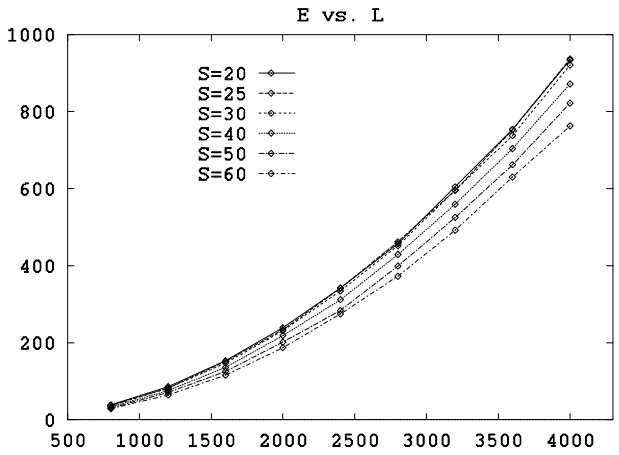

Figure 3.10: The maximal number of episodes, E (y-axis), stored to a criterion accuracy of approximately 97%, as a function of the number of L2 cells, L (x-axis), across variation in S parameter, but holding STC = 600 constant.

It is interesting to note that the percentage of increased horizontal weights, *H*, can reach such high levels as *S* increases. For example, when *S = 60*, 77.7% of all the horizontal weights have been set to 1. Figure 3.11, shows, for the *discrete correlograph* (Willshaw et al., 1969), the relationship between the total information I_W stored in the net and the fraction, *H*, of increased synapses. Maximal storage is achieved when exactly half of the synapses have been increased. Given the essential similarity of the basic storage and recall principles between the spatiotemporal model, TEMECOR, and the spatial discrete correlograph, it is presumably the case that I_W is maximal for topmost curve in Figure 3.10—i.e., *H = 50.0%, S = 20*.

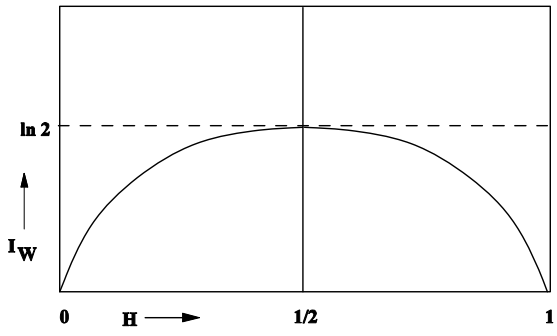

Figure 3.11: The relationship between the total information I_W stored in the net and the fraction, H, of increased synapses, for the discrete correlograph. From Nadal & Toulouse (1990).

Figures 3.12 and 3.13 show that the same qualitative relationships between *V* and *K* and between Y and *L*, respectively, hold across variation in *S*.

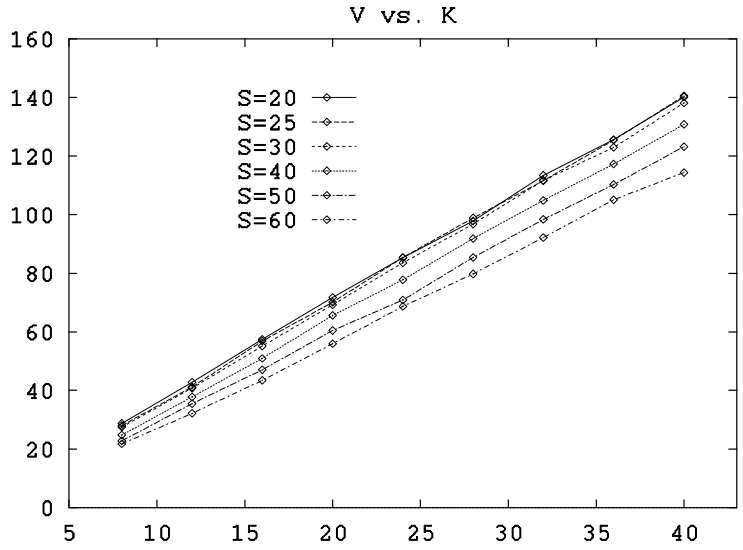

Figure 3.12: The linear relationship between V (y-axis) and K (x-axis) is preserved across variation in S. These eight curves correspond to the eight curves in the figure plotting E vs. L at the beginning of this subsection.

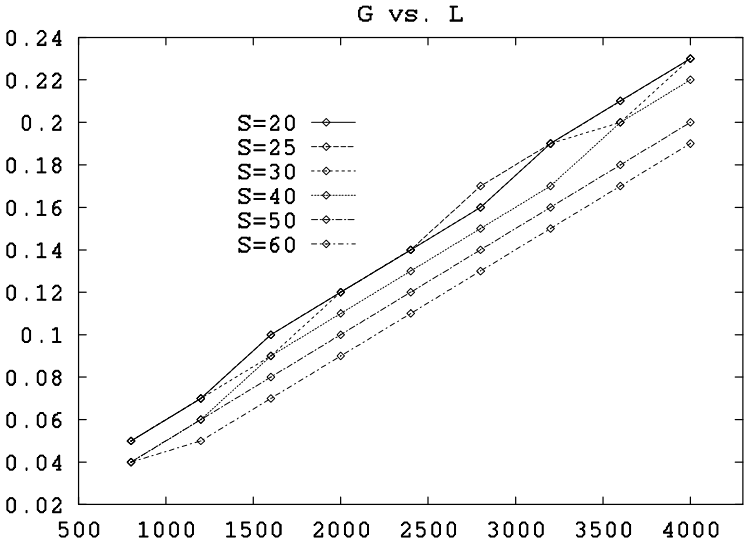

Figure 3.13: The linear relationship between Y (y-axis) and L (x-axis) is preserved across variation in S. As in the previous figure, these eight curves correspond to the eight curves in the figure plotting E vs. L at the beginning of this subsection.

3.6.2 Variation of γ parameter

The simulations reported in this section show how TEMECOR's capacity depends on the extent of horizontal connectivity, γ, between the cells of distinct CMs. All four curves in Figure 3.14 correspond to simulations involving uncorrelated episodes having *S = 20* and *T = 10*. Qualitatively similar behavior is expected for the case of correlated patterns. The figure shows that the rate drops off rather quickly as lower values of γ are considered. This is a problematic point for the model in terms of its neural plausibility since the degree of mono-synaptic, intrinsic interconnectivity of the neocortex, viewed even over regions encompassing on the order of tens of microcolumns, is probably far lower than γ = 0.7. The recurrent collateral plexus of CA3, which is considered to have one of the highest degrees of interconnectivity in the brain, only has a degree of connectivity of about 4.3% (Rolls, 1990). Nevertheless, the qualitative *faster-than-linear* relationship apparently holds across variation in γ. Figures 3.15 and 3.16 show the corresponding 'V vs. K' and 'Y vs. L' curves.

The sharp fall-off shown in Figure 3.14 is due to the fact that as the rate of connectivity falls below $\gamma = 1.0$, the distribution in the number of increased weights onto winning cells widens. Consider the learning that takes place between two successive L2 codes, Δ_t and Δ_{t+1}, both of size *S* and which have no CMs in common. When $\gamma = 1.0$, all members of Δ_{t+1} will have S weights increased. If Δ_t is later reinstated, then all cells in Δ_{t+1} will again have total input equal to S. Thus, θ can be safely set as high as S without causing any deletions in Δ_{t+1}.

In contrast, if $\gamma < 1.0$ then the expected total number of increases to any cell in Δ_{t+1} is $\gamma \times S$. Thus, if Δ_t is later reinstated, then the expected total input to the cells of Δ_{t+1} will equal $\gamma \times S$. However, setting $\theta = \gamma \times S$ will cause about half of the members of Δ_{t+1} to fail to become active. Thus, in order to avoid too many deletion errors, θ must be lowered below the expected total input value. The problem is that if θ is lowered too much, intrusion errors, which increase as a function of the number of stored associations, begin to accrue.

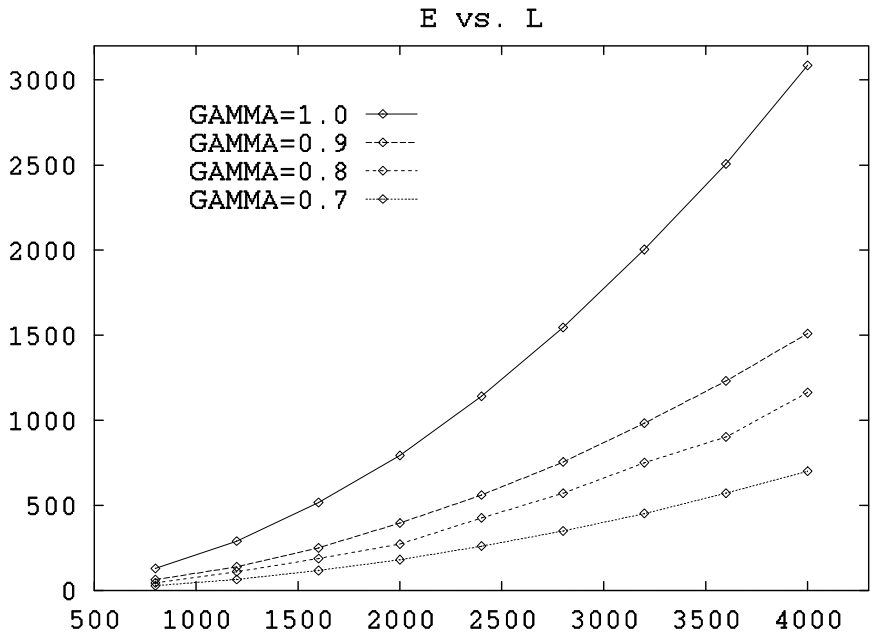

Figure 3.14: The rate of increase of E (y-axis) with L (x-axis) drops off rather quickly as the degree of horizontal connectivity is reduced; however, the qualitative faster-than-linear relationship is apparently preserved.

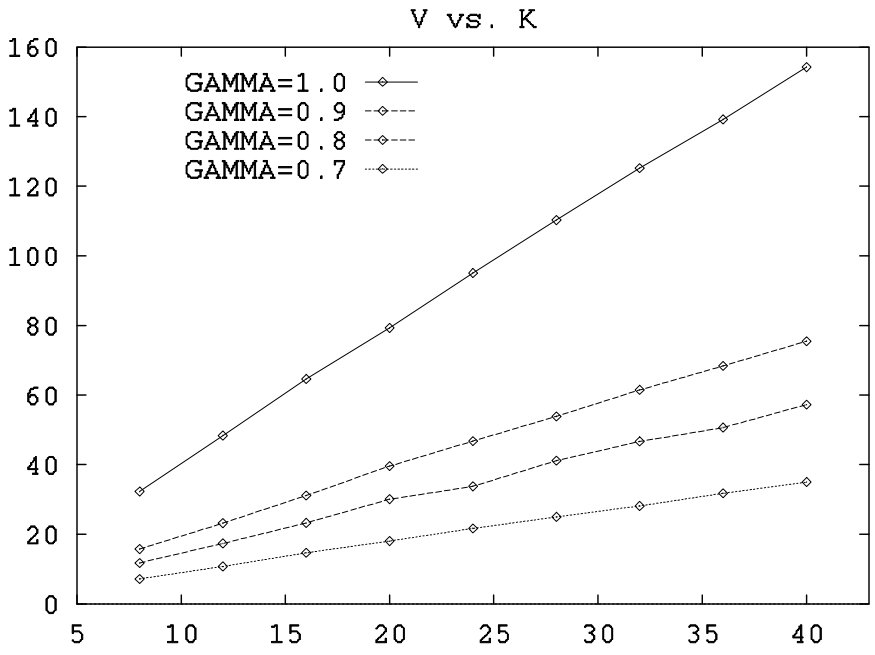

Figure 3.15: V vs. K curves corresponding to curves of previous figure.

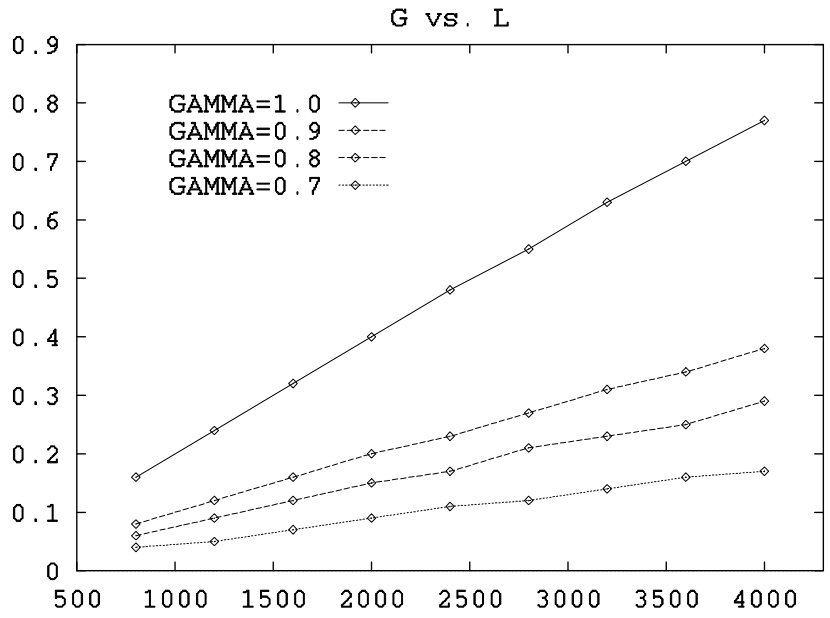

Figure 3.16: Υ ***vs. L curves corresponding to the curves in the two previous figures.***

3.6.3 Variation across desired recall accuracy

The simulations reported in this section show that the capacity barely increases at all even as the criterion recall accuracy is lowered substantially—i.e., from 97% to 75%. All simulations involved uncorrelated episodes having $S = 20$ and $T = 10$.

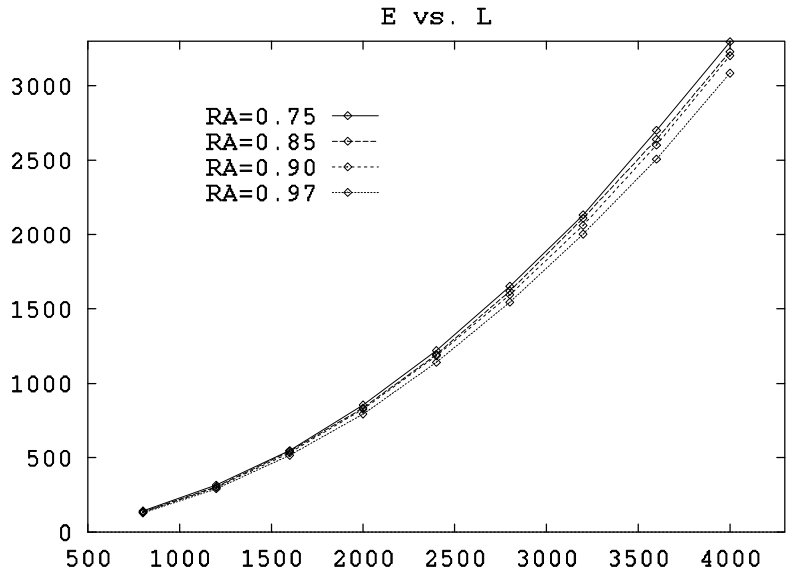

Figure 3.17: There isn't much capacity to be gained by reducing the desired recall accuracy from 97%, which was used in all other simulations reported herein, to values as low as 75%.

3.6.4 Highly redundant set of CSSs

Jordan (1986, p. 27) explains that repeated states are quite difficult for his model. Even a state sequence as simple as [ACABAA] takes 134 trials in order to learn. A *single* TEMECOR network, on the other hand, readily learned the set of complex state sequences in Table 3.4 given a single presentation of each. Each sequence consists of 20 state instances chosen from a set of only 4 unique states—A, B, C, and D. Each state consists of 25 (out of 100) features chosen at random. Other relevant parameters for this simulation were: CM size is 16, and $\theta = 21$. The overall recall accuracy for this simulation was 99.98%, which corresponded to only *two* errors (which were both intrusions) at the feature level (not the state level), out of 10,000 featural occurrences (i.e., 20 sequences × 20 states × 25 features/state). The percentage of increased horizontal weights was 8.2%. Note that this network was not full to capacity since θ was only 21. It could be moved up as high as 24 (since $S = 25$), thus screening out the intrusions while incurring no deletions. Thus, more sequences could have been stored. An intrusion error is one in which a feature (i.e., an L2 cell representing the feature) becomes active during the recalled trace when it was not active in the original trace.

3.6.5 Very long common subsequence

This set of simulations explicitly shows TEMECOR's capacity for handling sequences which have long runs of a single state in common (i.e., the 'tight loops' case of Smith & Zipser (1989). In particular, the two episodes (i.e., state sequences) used in this simulation are:

Γ^1: [B A A A A A A A A A A A A A A A A A A A D]

Γ^2: [C A A A A A A A A A A A A A A A A A A A E]

Episodes Γ^1 and Γ^2 have a 20-state long common subsequence of the state A. Here again, $S = 20$ out of a total of $M = 100$ features were chosen at random for each of the $U = 5$ states. Table 3.5 shows the results of several simulations for this data set.

Table 3.4: This is the list of 20 CSSs, each having 20 states, where the entire alphabet consisted of only U = 4 states.

Γ^{1}: [C C D A D B D C A B A C A B D B B B A A]

Γ^{2}: [A A A A B C C C C C A A A D B A A A C A]

Γ^{3}: [B C C B D B C B D C B D C B D D A B C C]

Γ^{4}: [A C D A B A A C A C C B A B C A C A B B]

Γ^{5}: [C C A B A D A A B C C B A B B C B C A B]

Γ^{6}: [A C C B D A B C A C D D A A A A D A A A]

Γ^{7}: [D D B D A D B C B B D B A C C D C D B D]

Γ^{8}: [B B C C B C C A C D B B C C B C B C A C]

Γ^{9}: [A A A A C C D A C B D D C B B D D A D C]

Γ^{10}: [D A D B D A D A D A D D A D D C B C C D]

Γ^{11}: [D C D B D A A D A A B D A A A D B A A A]

Γ^{12}: [A D D A C C C D A A D A C C B C C C B D]

Γ^{13}: [C C C C C D C D D C D D B C D D A B C B]

Γ^{14}: [B B B A D C C C A D B C B D B D C D D B]

Γ^{15}: [B C A C B D B B A D C C B D C A C A C C]

Γ^{16}: [D D B A D C D B B C D C B D A C D B D D]

Γ^{17}: [A C B B B D C D D A C C A D C D B C A C]

Γ^{18}: [A D A B D A D C B D B B B D D C C B C C]

Γ^{19}: [B A A B D D A A B B C B A C D D C C B D]

Γ^{20}: [B D A C C D B A A D C C D D D C C A C D]

Table 3.5: The results of several CSS simulations involving very long common subsequences of a single repeating state. Column abbreviations are as in earlier tables except that $\hat{V}$ = average number of times each L2 cell is used, but only for CMs corresponding to the features of state A.

E	θ	***K***	***L***	$\hat{V}$	R_{set}***(%)***	***H(%)***
2	18.0	8	800	5.0	99.5	1.93
2	16.0	8	800	5.0	99.1	1.93
2	14.0	8	800	5.0	97.8	1.93
2	13.0	8	800	5.0	96.6	1.93
2	18.0	12	1200	3.33	100.0	1.0
2	15.0	12	1200	3.33	100.0	1.0
2	12.0	12	1200	3.33	99.7	1.0
2	11.0	12	1200	3.33	99.2	1.0

These simulations show that either perfect or virtually perfect recall of these highly overlapped sequences is possible for a range of parameters. Specifically, two different size nets were used (*L = 800* and *L = 1200*) and a wide range of θ is tested in each case. Given that simulation six had 100% recall accuracy, that *S = 20* and that θ *= 15.0*, it follows that each of the 40 L2 codes corresponding to the 40 instances of state A must be different, in at least five CMs, from the other 39 L2 codes. In fact, in accord with Eq. 3.2, the expected overlap between any two of the L2 codes for A is 20/8 for the case of *L = 800* and 20/12 for the case of *L = 1200*. This performance in the 'tight loop' case contrasts strongly with that of the RTRL, which fails even for a sequence of five repetitions of the same state (Smith & Zipser, 1989).

The simulation results, for both uncorrelated and correlated patterns, show that TEMECOR-I scales well with problem size. Specifically, in both cases, the number of spatiotemporal patterns (episodes) that can be stored to criterion accuracy increases faster-than-linearly with the number of cells in the network. This finding is especially encouraging in the case of correlated patterns—i.e., complex state sequences (CSSs)—since, as stated in the introduction, linguistic information (phonemic transcriptions of utterances, for example) can be represented as sets of CSSs over a finite alphabet. Notice that in the last simulation (*K = 40*) of Table 3.3, the average number of instances of each state is 268.6. More importantly, the trend in $\hat{Q}$ is at least linear in *K*.

Furthermore, TEMECOR-I requires only a single presentation of each episode. It is also the case that since synaptic weights do not decrease, the memory traces of the episodes remain stable up to the point at which weight saturation effects lead to intrusion errors. Thus, even if a particular word is not accessed for an arbitrarily long period during which all the other words are accessed frequently, that word's trace will still read out perfectly when it finally is re-accessed. In contrast, models based on Backpropagation have been shown to be subject to massive ('catastrophic') forgetting (McCloskey & Cohen, 1989) in which newly encountered patterns obliterate old memory traces.

Chapter 4. The Enhanced Model: TEMECOR-II

4.1 Introduction

TEMECOR-I can be summarized as an unsupervised, distributed memory model, possessing some of the more general architectural and dynamical properties of the cortex of the mammalian brain, and whose storage capacity for non-orthogonal spatiotemporal feature patterns—and, as a special case, complex state sequences (CSSs)—scales faster-than-linearly with the size of the network, and which requires only a single presentation of each input. However, the model requires the unrealistic assumption, when doing recall testing, that prompting takes place at L2, rather than L1. This is analogous to the environment bypassing the sensory input pathways and directly activating internal representations (IRs) in deep cortices. Furthermore, because of the random method for choosing IRs (L2 codes), which depends neither on the previously active L2 code nor on the currently active input (L1 code), there is no way to access the correct initial IR. Also because of the random method of choosing IRs, the model fails to exhibit the property of continuity either in the mapping from L1 codes to L2 codes, via the F-projection, or in the mapping from L2 codes to (subsequent) L2 codes, via the H-projection. In other words, TEMECOR-I does not have the property that similar inputs map to similar IRs. Therefore the model fails to exhibit similarity-based generalization and categorization, which are the basis of many of those phenomena classed as semantic memory. Development of an integrated, general solution to these two problems has resulted in TEMECOR-II, which while more complex than its predecessor, has greater neural plausibility, and a much wider explanatory range for temporal sequence memory phenomena.

The property of continuity has been added to TEMECOR-II by grading the use of randomness in the process of mapping inputs to IRs. More specifically, continuity results if the IR chosen for an input is randomly changed by an amount that is inversely proportional to the similarity of that input and the set of previously experienced inputs. The basic idea is illustrated in Figure. 4.1. Suppose that the mapping between an input, A^1, and an IR, B^1, has been learned previously. The solid lines connecting A^1 to B^1 denote the increased connections (weights). Panel b shows another input, A^2,

having substantial overlap—i.e., similarity—with A^1, which results in strong, albeit sub-maximal, input to the cells comprising B^1 (shaded dark gray to reflect this level of support). We will refer to the set of IR cells receiving the highest amount of input as the *most-highly-implicated* IR. Now suppose that due to the sub-maximal level of support, a small amount of noise is added into the final selection of cells to become active at layer B (panel c), resulting in a final IR, B^2, slightly different from, although still substantially overlapping, B^1; specifically, $\left|B^1 \cap B^2\right| = 2$. The new learning that would occur in this case is depicted with dashed lines in panel c. Panel d shows another input, A^3, having a smaller overlap with A^1, reflected in the light gray shading of cells comprising B^1. Since the similarity between A^3 and A^1 is less than that between A^2 and A^1, relatively more noise is added into the process of choosing the IR, yielding a B^3 having smaller overlap with B^1 than does B^2; $\left|B^1 \cap B^3\right| = 1$. In the limiting case in which the current input has no overlap with— i.e., any similarity to— the previous inputs, the IR choice process becomes a completely random process, resulting in the minimal expected overlap between the resulting IR and the set of pre-existing IRs. Thus, the mapping exhibits continuity. We refer to the event in which a final winner—i.e., one resulting after noise has been added—is not one of the most-highly-implicated cells as an instance of *winner-flip*.

Note that although we use the terminology that ``noise is added" into the process of choosing internal representations, we can equivalently hypothesize that a certain baseline of noise is always present in the system and it is the deterministic quantities in the model—i.e., weights, activation levels, and thresholds—whose relative influence is actually directly modulated. This latter view is probably closer to reality but for convenience, we present the theory using the former terminology.

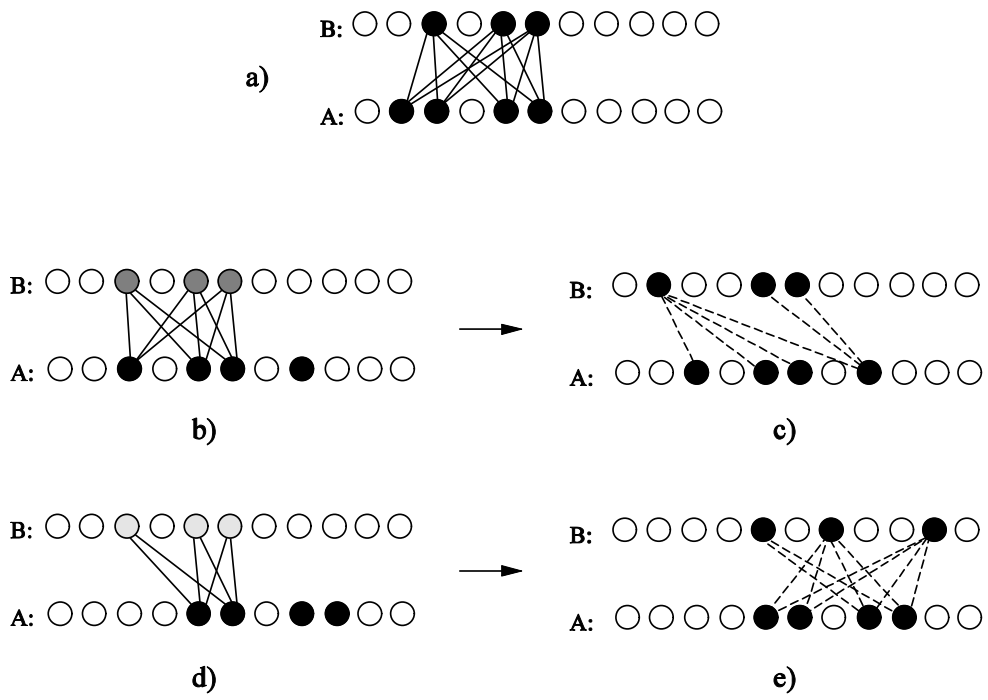

Figure 4.1: Illustration of the basic principle, used in TEMECOR-II, whereby addition of an amount of noise, inversely proportional to the similarity of current input to the set of previously learned inputs, results in a final mapping having the property of continuity. a) a pre-existing learned mapping between A^1 and B^1. b) Another input, A^2, highly similar to A^1. A relatively small amount of noise is added into the final choice of B^2 which thus, has high overlap with B^1, as seen in panel c. d) Another input, A^3, much less similar to A^1. A larger amount of noise is added into the winner selection process, resulting in a B^3 having smaller overlap with B^1 than does B^2. See text for more explanation.

Another important principle can be seen in Figure 4.1. Recalling that the most-highly-implicated IR is formally a set of L2 cells, the principle is that the membership of the most-highly-implicated IR generally remains invariant as the overlap between the current input and the previous input associated with the most-highly-implicated IR falls. In fact, in the simple example of Figure 4.1, we see that the most-highly-implicated IR is the same for both A^2, which has three out of four cells in common with A^1, and A^3, which has only two out of four cells in common with A^1. While the total input to the individual B cells in the most-highly-implicated IR decreases in going from the case of A^1 to A^2 to A^3, the membership (identity) of the most-highly-implicated IR remains the same. Thus, signals propagating via the plastic projection from layer A to layer B, would, without

other mechanisms, tend to selectively reactivate the internal representation associated with the most closely matching input, *even if the degree of match is low*. In this example, if we simply let the most-highly-implicated set of cells become active, then both inputs, A^2 and A^3, will be co-categorized with A^1 and the model will not be able to distinguish any of the A patterns. This condition can be described as *interference* between the memory traces.

Thus, the similarity-contingent addition of noise to the winner selection process not only achieves continuity, but also mitigates interference between traces, thus augmenting episodic storage capacity—i.e., storage capacity for the specific details of individual exemplars. Episodic information corresponds to the low order correlational information in the input set. It must be emphasized, however, that continuity and episodic capacity are conflicting goals. In terms of Figure 4.1, in order to maximize episodic capacity, the optimal strategy is to choose B patterns completely at random; that is, with no dependence on the plastic-matrix-filtered inputs. This maximizes the separation over the set of memory traces and, in conjunction with the other mechanisms and assumptions typically present in associative memory models (Willshaw et al. 1969; Amit, 1988; Anderson, 1972; Anderson, 1983; Hopfield, 1984; Kohonen, 1972, 1988; Moll et al., 1993; Rinkus, 1995), maximizes capacity. However, as discussed above, this minimizes the continuity of the mapping. Continuity increases with degree of dependence on the plastic-matrix-filtered inputs. O'Reilly & McClelland (1994) also describe and analyze this tradeoff between 'pattern separation' (i.e., capacity) and 'pattern completion' (which derives from the more basic property, continuity), and propose a theory in which the hippocampus minimizes the tradeoff. In particular, they view the mossy fiber projection from dentate gyrus (DG) to CA3 as a pattern separating transform which is engaged only when storing new information so as to maximize the separation between the newly formed trace and preexisting traces. When information is being retrieved, the DG-CA3 transform is not engaged, thus allowing the pattern completion properties of the CA3 to dominate [see Rolls (1990) for a similar viewpoint].

Hasselmo, Anderson & Bower (1991) and Hasselmo et al. (1992) provide neurophysiological evidence for another mechanism—one that is completely compatible with the noise-based mechanism proposed herein—for mitigating interference between traces. That work and its relation to TEMECOR-II are discussed in Sec. 4.11.

While Figure 4.1 illustrates the essential principle of using a similarity-contingent amount of noise to achieve continuity, it must be noted that the example is couched in terms of a purely

spatial mapping. In the TEMECOR-II model itself, the similarity measure takes the preceding temporal context—i.e., the series of states leading up to the current input—into account. The dependence of the length of the temporal window of sensitivity of the similarity metric on model parameters and on the statistics of the specific set of inputs experienced is a subject of future study. Note however, that the length of this temporal window, trailing the current input, within which past inputs can influence the current winner selection process is not a hard-wired, structural parameter of the model, as for example, is the case with the TDNN model of Waibel (1989). Furthermore, the extremely long contextual window of TEMECOR-I, as evidenced in the simulation results of Secs. 3.6.4 and 3.6.5, suggests that TEMECOR-II may also have such a long window of context.

The actual spatiotemporal similarity computation is formally divided into several stages as described in Sec. 4.5. Generally speaking, TEMECOR-II can be thought of as comparing its expectation as to what the current input should be with the actual current input. The model's expected input is manifest as the most-highly-implicated IR. On non-episode-initial time slices, the most-highly-implicated IR depends on the total pattern of synaptic inputs, via both the F-projection and the H-projection, to the L2 cells. The greater the match between the pattern of F-signals and the pattern of H-signals, the less noise added into the winner selection process, and the greater the extent to which the most-highly-implicated IR is reactivated. The lesser the match, the more noise added, and the less overlap between the newly formed trace and the set of previously embedded traces. These general contingencies entail the additional property, desirable from an information-theoretic standpoint, that the more familiar the input is (given the preceding context), and therefore, the less information present in it, the less learning that obtains. This general trend is illustrated in Figure 4.1. The number of newly increased synapses is larger for A^3 (panel e) than for A^2 (panel c), and is thus correlated with similarity to the previous input, A^1.

Various other models (Grossberg, 1976; Carpenter & Grossberg, 1987; Levy, 1989) also centrally involve a match process between an expected input and the actual input, and the hippocampus is considered to be the site at which the degree of novelty computation occurs (Hasselmo & Stern, 1995; Hasselmo et al. 1995; Levy, 1989; Carpenter & Grossberg, 1993; Gabriel et al. 1986).

Episode-initial time slices are automatically processed differently than non-episode-initial slices since, by definition, there is no previous temporal context—i.e., the vector of H-signals is the zero

vector. Accordingly, on episode-initial slices, the match computation collapses to the purely spatial case, measuring the similarity of the current input slice to all previously stored single time-slices.

The novelty computation provides a basis for moving back and forth between learning and recall modes. More to the point, it makes the distinction between learning and recall modes epiphenomenal on all but the time scale of the internal match computation. Thus, under TEMECOR-II, no explicit signal telling the model whether the current trial is a learning or recall trial is necessary. It simply is in learning mode if high noise is present (because high novelty has been detected), and is in recall mode if low noise is present (because high familiarity has been detected). More precisely, the instantaneous degree of match positions the model's instantaneous processing dynamics somewhere along a continuum between the two endpoints corresponding to pure recall and pure learning.

4.2 TEMECOR-II's Various Processing Modes

TEMECOR-II has two primary processing modes differentiated by whether or not the vector of F-signals—i.e., input from the world—is suppressed. Suppression has two consequences on TEMECOR-II's operation: a) the novelty detection stage which matches H-signals against F-signals can no longer function, and b) there is no feedforward input to L2. This mode in which the outside world is effectively shut off is referred to as the *solipsistic mode*. It is necessary in order to accommodate situations in which humans are lost deep in thought, or imagination, or reminiscence. There can still be noise present in solipsistic mode. However, it cannot be under the control of the degree of match between expected and actual input because the actual input is turned off. Thus, some other source of noise control must be posited for solipsistic mode. This issue is not pursued herein.

The other mode, in which the F-signals are taken into account, is called the *interactive mode*. This is the mode that all of the preceding explanation in this chapter has concerned.

The interactive vs. solipsistic distinction (and this too is most likely a continuum) is orthogonal to the learning vs. recall distinction. Thus each of the two primary modes has one sub-mode in which new learning occurs and one characterized primarily by the read-out of pre-existing information. It must be emphasized that the primary distinction—solipsistic vs. interactive—is

controlled by a variable that is presently not modeled within the theory.[21], whereas the secondary distinction—learning vs. recall—is, as explained above, purely a function of the perceived novelty of the current situation which has a purely mechanistic explanation within the theory.

The '2 x 2' distinction leads to four different operational sub-modes which we now name and describe in phenomenological terms. Suppose the model is operating in interactive mode and the currently unfolding episode is completely familiar. If the F-input is then suddenly cut off, the model should be able to complete the read out of the internal trace—i.e., remember how the episode finishes. During the period in which the internal trace is continuing to be read out even though external inputs have been shut off, the model is said to be in *solipsistic recall* or *reminiscence mode*.

If we now imagine that significant noise is introduced while in solipsistic mode, then, even if the currently unfolding L2 trace originated from a familiar situation, novel L2 swaths of activity will obtain, but they will not be influenced by the current real world (i.e., F) input. Since novel L2 traces will be occurring, learning will take place. We refer to this as *solipsistic learning* mode. This mode offers a possible basis for the psychological states of fantasizing and dreaming. Note that the novel L2 traces that arise in this mode are not completely random. This is due to the deterministic influence of the H-signals. This mode is not explained further herein but is described to provide a context for the other modes.

If the model is in interactive mode and the currently unfolding situation is perceived as familiar, then as stated earlier, the trace that obtained while experiencing the original similar instance will be reactivated to a great extent, thus little or no new learning will occur. This is not properly viewed as 'recall' mode, since after all, the actual episode to be remembered is presently recurring. Thus we refer to this mode as *interactive tracking* mode. This corresponds to those times when a human just passively tracks (i.e., witnesses) the unfolding of a highly familiar event.

Finally, if the model is in interactive mode and the currently unfolding episode is unfamiliar, as stated earlier, the newly formed trace will be highly distinct from any pre-existing traces, thus much learning will take place. We refer to this as *interactive learning* mode.

---

[21] Presumably, whether the model is in solipsistic mode vs. interactive mode is at least partly due to affective signals, but this is not explicitly addressed herein.

4.3 Properties for TEMECOR-II

4.3.1 Property 1: Uses L1 codes as prompts, not L2 codes

As stated in the first paragraph of this chapter, one of the main problems with TEMECOR-I is that it requires the use of L2 codes as prompts rather than L1 codes. L1 is TEMECOR-II's interface with the world. Therefore, recall prompts, whether they be single time slices or whole sequences, should be L1 codes. TEMECOR-II exhibits this property and it is demonstrated in all of the simulations.

4.3.2 Property 2: Spatiotemporal generalization

The model exhibits continuity in the spatiotemporal domain. That is, the similarity between the internal representations increases as a function of the similarity of the corresponding inputs. As explained in Sec. 4.1, this property:

a) is the more basic property supporting generalization and categorization, and
b) implies that in general, there will be less learning, in both the H- and F-projections, during familiar episodes than during novel ones.

This latter property is directly shown in Sec. 4.10.2 and is therefore taken as indirect evidence that the model has continuity and therefore generalization and categorization. Categorization capability will be shown more directly in the simulations of Sec. 4.10.5. We refer to these properties/capabilities generally as the *spatiotemporal generalization* property.

4.3.3 Property 3: Complex sequence disambiguation

Like its predecessor, TEMECOR-II is capable of remembering sets of complex state sequences (CSSs) on the basis of single-trial learning. We refer to this property as the *complex sequence disambiguation* property and it is demonstrated in the simulations of Secs. 4.10.3 and 4.10.4.

4.3.4 Property 4: Multiple competing hypotheses (MCH)

The model maintains multiple competing hypotheses (MCHs) when given ambiguous prompt information. As subsequent disambiguating information (i.e., successive states of the prompt) enters the system, the disconfirmed hypotheses fade away. Suppose the model has previously learned the following three episodes.

Γ^1: [A B C D E]

Γ^2: [G B C F K]

Γ^3: [L R B D M]

Now suppose that the model is prompted with the state, B. In this case, an L2 code approximating (or at least, containing significant portions of) the union of the L2 codes for state B Γ^1 and for state B in Γ^2 and for state Γ^3—i.e.,

$$\Delta^1_B \cup \Delta^2_B \cup \Delta^3_B$$

—should become active. Then, if the next state of the prompt is C, an L2 state of activity approximating

$$\Delta^1_C \cup \Delta^2_C$$

should become active. There should be no component of L2 activity due to Δ^3_D at this point because the hypothesis that the requested episode is Γ^3 has been ruled out. Finally, if the next state of the prompt is D, then Δ^1_D should become active and this should lead to Δ^1_E.

The question is: how precisely do we represent MCHs? There are at least two possibilities. We could allow something like the union of all the L2 codes corresponding to the hypotheses to become active simultaneously. In our example, when B presents, this implies that 3 x *Q* L2 cells, 3 cells per CM, would be co-active. This constitutes a deviation from the assumption that the CMs are winner-take-all modules and so is not developed herein. On the other hand, rather than letting 3*Q* L2 cells be active, we could instead randomly pick one of the three (by assumption, equally-

strongly-implicated) cells to become active in each CM. In this case, each of the competing hypotheses—*qua* L2 codes—is expected to win in $Q/3$ of the CMs. This latter method for representing MCHs is used herein and is described more fully in Sec. 4.5.3. The relevant simulation results are in Sec. 4.10.3.

4.4 Architecture of TEMECOR-II

Two essential architectural changes must be made to TEMECOR-I in order to accommodate all of the properties listed in the previous section.

a) The feedforward (F) projection from L1 to L2 must be generalized. As depicted in Figure 4.2, under TEMECOR-II, each L1 cell synapses upon all L2 cells. Furthermore, these F-weights are binary-valued and plastic. In fact, a *reciprocal* (R) weight is assumed for every F-weight. The R-weight is increased at the same moment as the corresponding F-weight. As in TEMECOR-I, these reciprocal projections are needed to explain how read-out of an L2 swath can cause the corresponding L1 swath to read-out along with it.
b) Auxiliary circuitry must be added which allows for the ongoing—i.e., time slice by time slice—computation of a) the degree of match between the model's expectancy and the actual input, and b) the attendant noise which is injected into the winner choice process. The whole TEMECOR-II model is depicted in Figure 4.3.

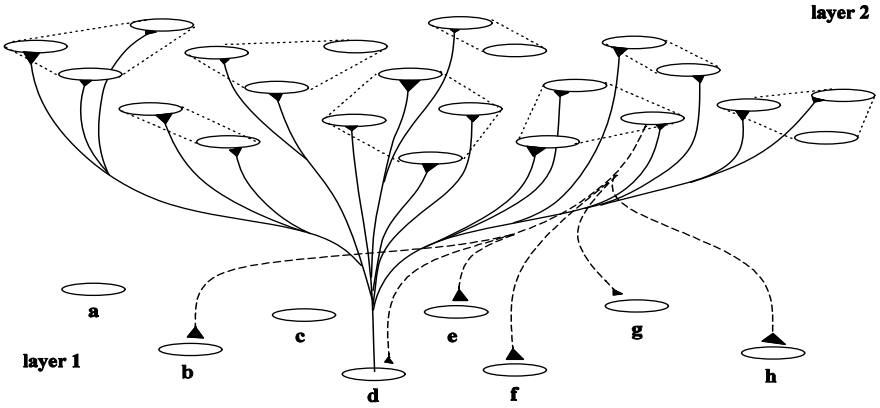

Figure 4.2: The generalized feedforward (F) and reciprocal (R) projections of TEMECOR-II. Only a few connections from one L1 cell are shown. Similarly, only a few R-connections from one L2 cells are depicted. Full connectivity is generally assumed for both the F- and R-projections herein.

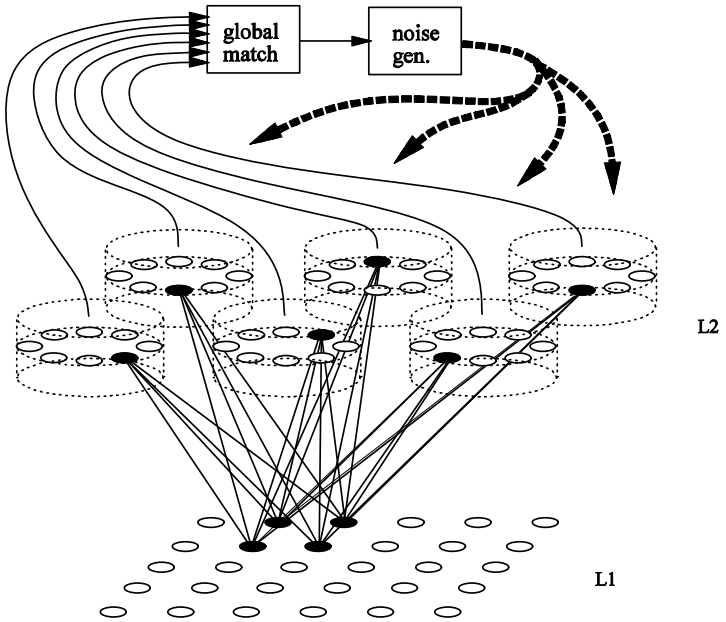

Figure 4.3: The whole TEMECOR-II model. The cylinders are intended to suggest the mini-columns of cortex. The lines leading to the global matching module carry the results of the local matching computations that take place within each CM. The global degree of match is then used to determine how much noise to inject into the winner selection process. Horizontal connections exist within L2 but are not depicted. The vertical projections between L1 and L2 are bi-directional and plastic. This figure shows a particular input, consisting of the four active (black) cells in L1 and the corresponding internal representation consisting of the six active L2 cells.

Change in interpretation of L2 under TEMECOR-II

Under TEMECOR-I, one L2 cell is chosen in each *active* CM. Thus if *S* L1 cells were active (i.e., *S* features present) on a given time slice, then *S* L2 cells would also be active. However, since there is no longer a 1-to-1 correspondence between L1 cells and L2 CMs under TEMECOR-II, the concept of an *active* CM must be modified. Two design choices are possible: a) allow all CMs to be active on every time slice, or b) choose a subset of CMs to be active. Method a was chosen because it is simpler than method b. Specifically, method b requires extra mechanism for deciding which subset of CMs will be active, whereas method a does not. Note that since an overarching concern is to achieve low coding rates, particularly in L2, and since method (a) has all CMs active,

method (a) also generally implies the use of larger CM sizes than method (b). Thus, under TEMECOR-II, all L2 codes will contain *Q* cells, where *Q* is the number of CMs in L2.

This amounts to a significant change in the interpretation of the L2 cells. Under TEMECOR-I, L2 cells formally had the exact same feedforward receptive fields as their corresponding L1 cells. The L2 cells could therefore be considered to be representing the same features as the L1 cells. But under TEMECOR-II, this is no longer the case. Now the L2 cells' feedforward receptive fields formally correspond to hyper-regions of the *M*-dimensional feature space, $\{0,1\}^M$, defined by L1. Thus under TEMECOR-II, the feedforward receptive fields of the L2 cells correspond to complex spatial configurations of features. Thus TEMECOR-II is most analogous to the deeper regions of neocortex—anterior inferotemporal (IT) and prefrontal. Evidence has already been cited in Sec. 3.5 for the general trend of increasing receptive field complexity as we move from earlier to later cortices.

In addition, the feedforward receptive fields of two L2 cells from the same CM can in general correspond to very different regions of hyperspace. Thus CMs can no longer be identified with L1 cells. A complete characterization of the L2 receptive fields must also take into account the modifiable horizontal connections. These render the L2 receptive fields formally spatiotemporal in nature and can, in principle, explain why certain cells fire in certain spatiotemporal contexts but not in others, even though the spatial inputs themselves (i.e., snapshots) might be highly similar in the two contexts.

4.5 TEMECOR-II's Processing Algorithm

The stages comprising the computational cycle that TEMECOR-II performs on every time slice are described in this section. Stages three, four and five are omitted on episode-initial time slices. Specifically, this algorithm corresponds to the *interactive processing mode* described in Sec. 4.2. It is assumed that all L1 codes—i.e., input patterns—have *S* active cells and that all L2 codes have *Q* active cells (recall that *Q* is the number of CMs in L2). Sec. 4.6 describes the solipsistic mode variant of the basic algorithm.

Table 4.1 provides definitions of various symbols and terms used in conjunction with TEMECOR-II. Some of the symbols are redefined from their usage in Ch. 3. As in Ch. 3 all of

these terms are also defined in the text as they are introduced. Note that the superscript *p* denoting the episode is sometimes dropped from the symbol definitions, as is the subscript, *t*.

Table 4.1: Table of definitions for symbols involved with TEMECOR-II

Symbol	Definition
Q	The number of CMs.
$\phi_t^p(x)$	The total synaptic input to cell, *x*, from the set of cells active on the previous time slice. $\phi_{i,t} = \sum_{j \in \Delta_{t-1}} w_{ji} \quad , t > 0$
$\hat{\phi}_t^p$	Vector of $\phi_t^p(x)$ values over all L2 cells
${}_i\breve{\phi}_t^p$	Max. total horizontal input to any cell in CM_i at time *t* of episode, *p*. $= \max_{x \in CM_i} \phi_t^p(x)$
$\Phi_t^p(x)$	Normalized ϕ value for cell, *x*, with respect to all the cells in *x*'s CM.
$\psi_t^p(x)$	Total synaptic input to cell, *x*, from the set of currently active L1 cells. $\psi_{i,t} = \sum_{j \in \Gamma_t} w_{ji}$
$\hat{\psi}_t^p$	Vector of $\psi_t^p(x)$ values over all L2 cells
${}_i\breve{\psi}_t^p$	Max. total vertical input (i.e., F-input) to any cell in CM_i at time *t* of episode, *p*. $= \max_{x \in CM_i} \psi_t^p(x)$
$\Psi_t^p(x)$	Normalized ψ value for cell, *x*, with respect to all the cells in *x*'s CM.
${}_i\sigma_t$	The set of cells in CM_i having the highest H-input $\left\lvert {}_i\sigma_t \right\rvert$ may be greater than one, i.e., there cane be ties.
σ_t^p	The set of most highly favored cells, on the basis of H-input, across all L2 cells. $= \bigcup_i \sigma_t$

${}_i\Omega_t$	The set of cells in CM_i having the highest F-input $\left\|{}_i\Omega_t\right\|$ may be greater than one, i.e., there cane be ties.
Ω_t^p	The set of most highly favored cells, on the basis of F-input, across all L2 cells. $=\bigcup_i \Omega_t$
$\chi_t^p(x)$	On non-episode-initial slices, it measures the degree of match between the horizontal inputs (mediating information about the prior temporal context leading up to the current input) and the feedforward inputs (mediating the current input), i.e., a spatiotemporal match measure. $=\Phi_t^p(x)^u \times \Psi_t^p(x)^v$ On episode-initial slices, it measures the degree of match between the current input and all previous inputs, i.e., a spatial match measure. $=\Psi_t^p(x)^w$
${}_i\breve{\chi}_t$	Maximum computed match to any cell in CM_i at time *t* of episode, *p*. $=\max_{x\in CM_i}\chi_t^p(x)$
$\mathrm{X}_t^p(x)$	Normalized χ value for cell, *x*, with respect to all the cells in *x*'s CM.
${}^H\Theta$	Horizontal activation threshold
${}^F\Theta$	Feedforward (bottom-up) activation threshold
${}^R\Theta$	Reciprocal (top-down) activation threshold
${}^\chi\Theta$	Threshold above which a 100% match is considered to exist.
${}_i\pi$	Maximum X value in CM_i.
G	Overall, i.e., averaged over all CMs, measure of match between expected and actual input. $G=\sum_{i=1}^{Q}{}_i\pi \Big/ Q$
${}_R\nu_t$	The range of ν values to which X values will be mapped at *t*.
$\nu_t^p(x)$	An intermediate variable used to determine the final probability that cell, *x*, will be chosen winner in its CM at time, *t*.

$\rho_t^p(x)$	The final probability that cell, *x*, will be chosen winner in its CM at *t*.
$\xi_t(x)$	The number of cells, x, in CMi, for which $\chi_t(x)=_i\breve{\chi}_t$.
Ξ_t	The number of multiple competing hypotheses (MCHs) active at *t*.
η_t^p	Learning rate parameter at time, *t*, of episode, *p*.
$\varsigma_t(x)$	The total R-input to L1 cell, *x*, from the active L2 cells. $=\sum_{j\in\Delta_t} w_{jx}$

4.5.1 Compute the total feedforward input for each L2 cell

The total input via the feedforward projection, $\psi_t(x)$, to an L2 cell, *x*, at time, *t*, is:

$$\psi_{i,t} = \sum_{j\in\Gamma_t} w_{ji} \tag{4.1}$$

where Γ_t is the set of active L1 cells on the current time slice and w_{jx} is the weight from cell *j* in Γ_t to L2 cell, *x*. $\hat{\psi}_t$ is the entire vector of $\psi_t(x)$ values, across all L2 cells, at *t*.

4.5.2 Compute the normalized feedforward inputs

Let $_i\breve{\psi}_t$ be the maximum ψ value across all cells in CM_i on time slice *t*.

$$_i\breve{\psi}_t = \max_{x\in CM_i} \psi_t(x) \tag{4.2}$$

The normalized values are:

$$\Psi_t(x) = \frac{\psi_t(x)}{\max(_i\breve{\psi}_t, {}^F\Theta_t)} \tag{4.3}$$

The reasoning for Eq. 4.3 is as follows. If there is a cell, *x*, in the CM under consideration such that $\psi(x) \geq {}^{F}\theta$, then *x* is considered to be perfectly matched to the current feedforward input—that is, there is maximal evidence from the feedforward input that *x* should become active. Accordingly, $\Psi_t(x) = 1$. ${}^{H}\theta$ is related to the *vigilance* parameter of the ART models (Carpenter & Grossberg, 1987) in that it helps to determine the coarseness of categories formed by the model. However, how TEMECOR-II deals with spatiotemporal categories and category coarseness depends on other parameters besides ${}^{F}\theta$ and on the interactions between ${}^{F}\theta$ and these other parameters. In general, as ${}^{F}\theta$ is increased, categories become finer.

On the other hand, if no cell has $\psi \geq {}^{F}\theta$ then that indicates that no cell is maximally implicated by the current input. Accordingly, all Ψ values for cells in this CM will be less than one. There will still, in general, be a distribution of Ψ values reflecting the distribution of ψ values, but the maximal Ψ value will be less than one.

We refer to the set of L2 cells having maximal F-input in CM_i, at time *t*, as the *feedforward expectancy* (F-expectancy) in *i* at *t*, denoted as ${}_i\Omega_t$. Members of ${}_i\Omega_t$ may also be referred to as *F-winners*. Note that merely being an F-winner does not guarantee that the cell will ultimately be chosen to become active on the current time slice, since the final set of winners also depends on the H-inputs and noise. Also, note that $\left|{}_i\Omega_t\right|$ may be greater than one.

The overall F-expectancy, Ω_t, is the union of the F-expectancies at the individual CMs.

$$\Omega_t = \bigcup_i {}_i\Omega_t \tag{4.4}$$

4.5.3 Compute ${}^{H}\theta$

The model assumes that the horizontal activation threshold parameter, ${}^{H}\theta$, has, for any particular simulation, a global baseline value, ${}^{H}\theta_{baseline}$, which is set to achieve a particular balance between capacity and generalization capability. In general, the higher we set ${}^{H}\theta_{baseline}$, the greater the separation between memory traces and the greater the capacity. In contrast, as we lower the ${}^{H}\theta_{baseline}$, the more overlapped the memory traces, the better the generalization capability, and the coarser the resulting spatiotemporal categories formed by the model. Note that the practical upper limit for ${}^{H}\theta_{baseline}$ is *Q-1*. This stems from the facts that a) exactly *Q* L2 cells are active on every

time slice, and b) individual L2 cells do not synapse upon themselves. Thus, the highest ϕ value an L2 cell can ever have is *Q-1*. The practical lower limit for ${}^{H}\theta_{baseline}$ is simulation-specific and depends on the particular statistics of the set of episodes presented as well as other model parameters.

The instantaneous value, ${}^{H}\theta$, may generally deviate from the baseline value depending on the immediate situation. In particular, given the means by which multiple competing hypotheses (MCHs) are represented (described in Sec. 4.3.4), ${}^{H}\theta$ must be scaled downward on time slices, *t*, by an amount proportional to the number of MCHs, Ξ_{t-1} that existed at *t-1*. Accordingly,

$$ {}^{H}\theta_{t} = \frac{{}^{H}\theta_{baseline}}{\Xi_{t-1}} \tag{4.5} $$

This is necessary because on time slices following time slices on which n MCHs existed, the maximal expected ϕ values will be approximately $\frac{Q-1}{n}$, rather than approximately *Q-1*. In order to register a perfect match at *t+1*, should one actually exist (i.e., should the input at *t+1* actually be one of the inputs expected on the basis of one of the MCHs at *t*), ${}^{H}\theta$ must not be greater than $\frac{Q-1}{n}$.

Due to the influence of ${}^{H}\theta_{baseline}$ and its derived parameter, ${}^{H}\theta$, upon the generalization and categorization properties of the model, these parameters, in conjunction with ${}^{F}\theta$, can be viewed as performing a function related to the *vigilance* parameter in the ART models of Carpenter & Grossberg (1987), except that, as pointed out in the previous section, this functionality is with respect to the spatiotemporal pattern domain.

4.5.4 Compute the total horizontal input for each L2 cell

The total input via the horizontal projection, $\phi_t(x)$, to an L2 cell, *x*, at time, *t*, is:

$$ \phi_t(x) = \sum_{j \in \Delta_{t-1}} w_{jx} \tag{4.6} $$

where Δ_{t-1} is the set of active L2 cells on the previous time slice. $\hat{\phi}_t$ is the entire vector of $\phi_t(x)$ values, across all L2 cells, at *t*.

4.5.5 Compute the normalized horizontal inputs

Let ${}_i\phi_t$ be the maximum ϕ value across all cells in CM_i on time slice *t*.

$$ {}_i\breve{\phi}_t = \max_{x \in CM_i} \phi_t(x) \tag{4.7}$$

The normalized ϕ values are:

$$ \Phi_t(x) = \frac{\phi_t(x)}{\max({}_i\breve{\phi}_t, {}^H\Theta_t)} \tag{4.8}$$

The reasoning for Eq. 4.8 is as follows. If there is a cell, *x*, in the CM under consideration such that $\phi(x) \geq {}^H\theta$, then *x* is considered to be perfectly matched to the current horizontal input—that is, there is maximal evidence from prior context that *x* should become active on the current time slice. Accordingly, $\Phi_t(x) = 1$. On the other hand, if no cell has $\phi \geq {}^H\theta$ then that indicates that no cell is maximally implicated by the prior context. Accordingly, all Φ values for cells in this CM will be less than one. There will still, in general, be a distribution of Φ values reflecting the distribution of ϕ values, but the maximal Φ value will be less than one.

We refer to the set of L2 cells having maximal H-input in CM_i, at time *t*, as the *horizontal expectancy* (H-expectancy) in *i* at *t*, denoted as ${}_i\sigma_t$. Members of ${}_i\sigma_t$ may also be referred to as *H-winners*. Note that merely being an H-winner does not guarantee that the cell will ultimately be chosen to become active on the current time slice, since the final set of winners also depends on the F-inputs and noise. Also, note that $\left|{}_i\sigma_t\right|$ may be greater than one.

The overall H-expectancy, σ_t, is the union of the H-expectancies at the individual CMs.

$$\sigma_t = \bigcup_i {}_i\sigma_t \tag{4.9}$$

4.5.6 Compute overall degree of match for each L2 cell

4.5.6.1 Episode-initial case

As there is no previous context on episode-initial time slices to compare with the current feedforward input vector, the degree of match is based purely on how closely the current spatial input matches the closest-matching previous spatial input. This can be computed directly at the L2 cells. For example, if there has only been one previous input and it has 10 out of $S = 20$ features in common with the current input[22], then in any given CM, the L2 cell, x, which was chosen winner for that previous input will now have 10 active inputs (i.e., a $\psi(x) = 10$), whereas the other L2 cells in that CM will have ψ values of zero. If we assume ${}^F\theta = 15$, then by Eq. 4.3, we have $\Psi(x) = 10/15 = 2/3$. As the Ψ values are constrained to the interval, [0,1], the exponent, w, in Eq. 4.10 provides a means for controlling the shape of the generalization gradient. An exponent parameter is used for similar purposes in the MINERVA 2 model (Hintzman, 1986). Figure 4.4 shows that increasing w sharpens the gradient, and at the same time makes the match criterion more stringent. Thus, in general, a higher w will lead to lower match values, more noise being added to the winner selection process, and less overlap over the set of stored memory traces.

$$\chi_t(x) = \Psi_t(x)^w \tag{4.10}$$

[22] Recall that all time slices are assumed to have the same number, S, of active features.

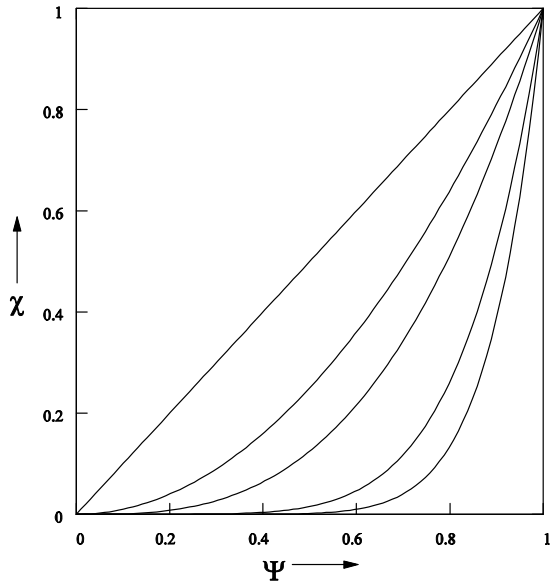

Figure 4.4: Various instances of the Match function (episode-initial case) corresponding to w values: 1, 2, 3, 6 and 9.

4.5.6.2 Non-episode-initial case

On non-episode-initial time slices, the overall degree of match, $\chi(x)$, for an L2 cell, x, depends on both the total normalized horizontal input, $\Phi(x)$ which represents the degree of support from prior context, and the total normalized feedforward input, $\Psi(x)$ which represents the degree of support from the current input. Again, the exponents (u and v in this case) control the strictness of the match. Specifically,

$$\chi_t(x) = \Phi_t(x)^u \Psi_t(x)^v \qquad (4.11)$$

where $u,v \geq 1$. Figures 4.5a,b depict the χ function for $u,v = 2$ and $u,v = 3$. Other types of functions—e.g. sigmoidal—will also be investigated in future work. Since $0 \geq \Phi(x) \geq 1$ and $0 \geq \Psi(x) \geq 1$, $0 \geq \chi(x) \geq 1$.

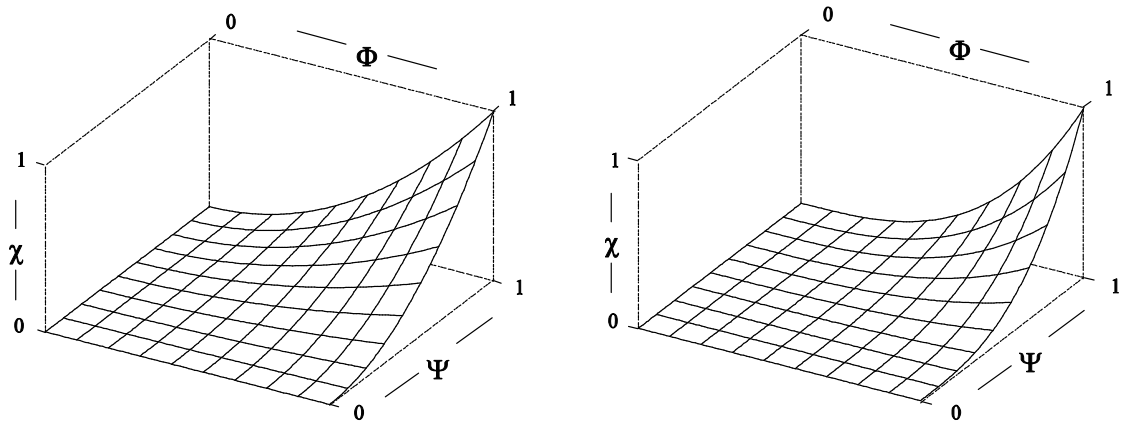

Figure 4.5: Match function (non-episode-initial case), χ, for the case of u,v = 2 (left) and for the case of u,v = 3 (right).

Figure 4.6 depicts the four qualitatively different ways in which a *high match condition* can exist in a given CM_i. The axons that descend from above onto the L2 cells in this and subsequent similar figures denote horizontal connections. In this figure and the next several similar figures, $S = 4$, $Q = 5$, ${}^{H}\theta = Q\text{-}1 = 4$, and ${}^{F}\theta = S$. Panel a portrays the case in which ${}_i\sigma$ contains two cells whereas the F-expectancy, ${}_i\Omega$, contains only the shaded cell, which we will refer to as cell *x*. In this case, $\chi(x) = 1.0$ and the χ value of the cell receiving only H-input would be 0.0. Panel b depicts an even simpler case in which ${}_i\sigma$ and ${}_i\Omega$ each contain exactly one cell and that cell is common to both; again $\chi(x) = 1.0$. Panel c shows the case where ${}_i\Omega$ is a superset of ${}_i\sigma$ and again, $\chi(x) = 1.0$. Finally, panel d depicts the case in which ${}_i$\mho and ${}_i\Omega$ intersect at more than one cell. This corresponds to the MCH case described earlier.

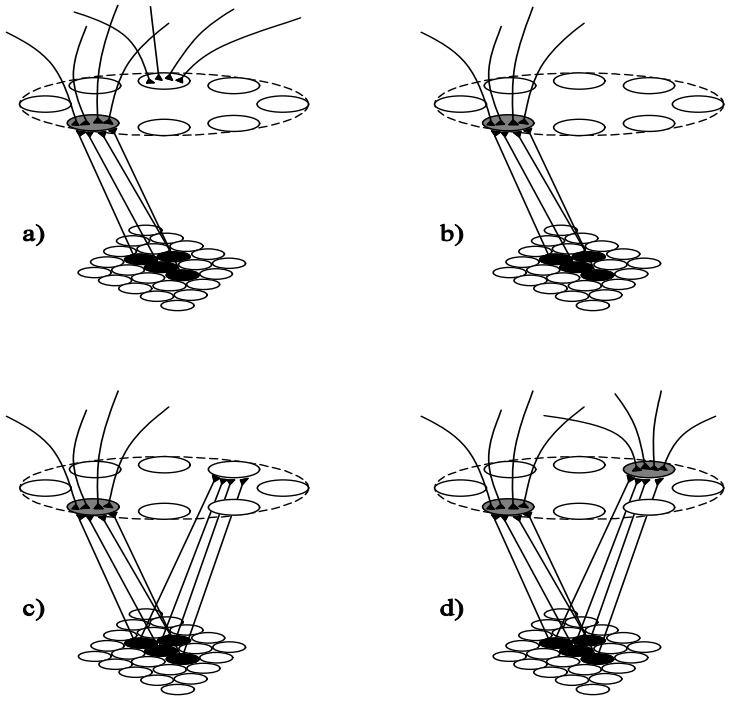

Figure 4.6: Four qualitative ways a high match condition can exist in a given CM. The connections impinging, from above, on the L2 cells correspond to horizontal connections from other L2 cells in other CMs.

Figure 4.7 leads to the same χ values as in Figure 4.6 except that it is more realistic because it includes spurious horizontal and vertical signals.

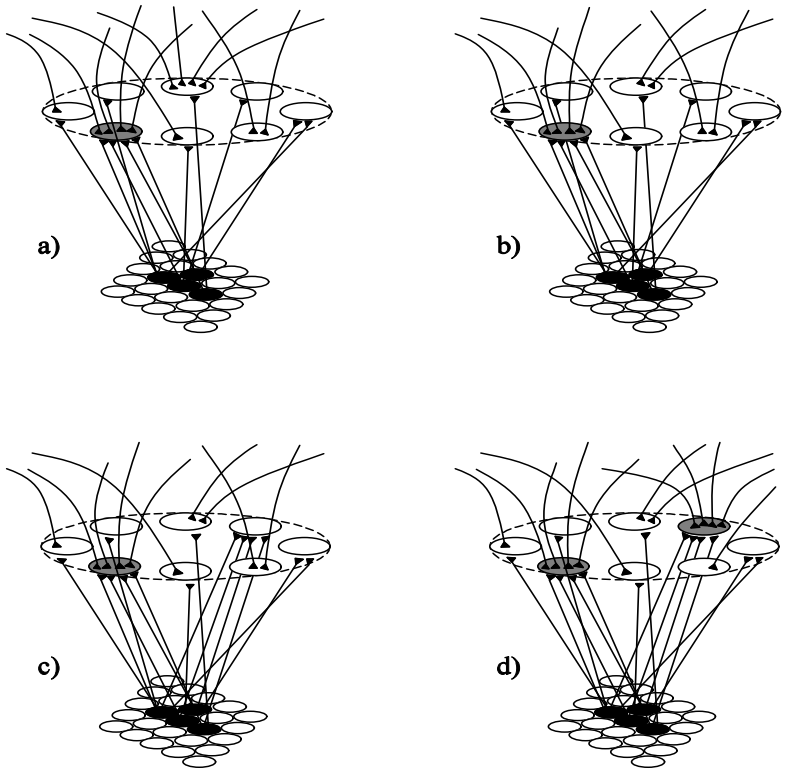

Figure 4.7: More realistic depiction of the four ways a high match condition can exist in a given CM. Many spurious signals exist but χ = 1.0 only for the shaded cells.

According to Eq. 4.11, $\chi(x)$ depends continuously on $\Psi(x)$. Figure 4.8 depicts match conditions of four different magnitudes based on different Ψ values for the shaded cell, *x*, assuming ${}^{H}\theta = 4$ and ${}^{F}\theta = 4$. Panel a is a repeat of Figure 4.6a; $\chi(x) = 1.0$. The match condition in panel b yields $\chi(x) = (0.75)^2 = 0.5625$. That of panel c yields $\chi(x) = (0.5)^2 = 0.25$ and that of panel d yields $\chi(x) = (0.25)^2 = 0.0625$.

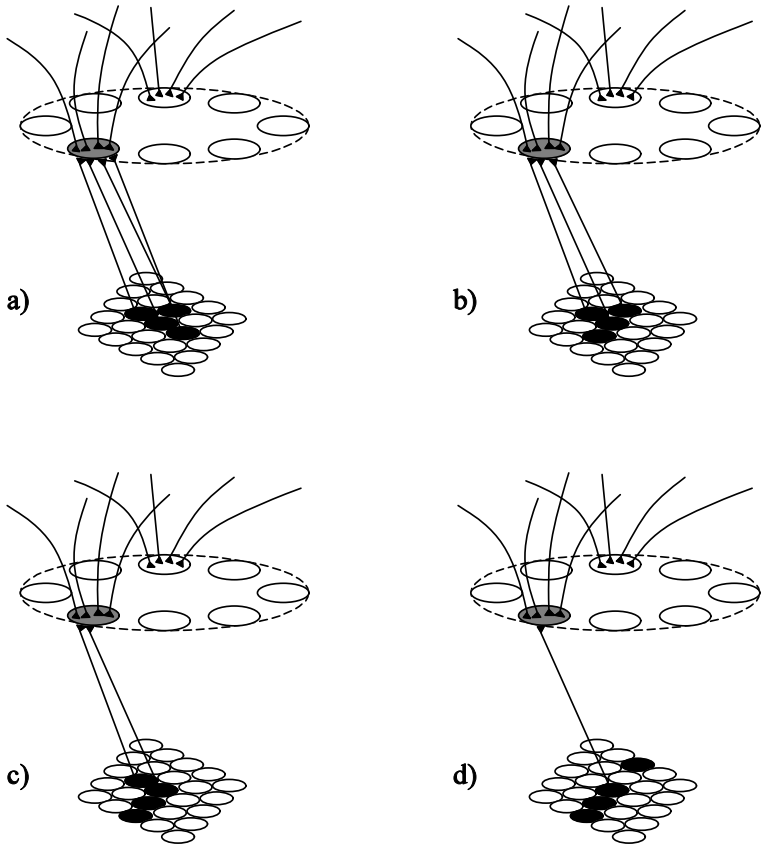

Figure 4.8: Match conditions of four different magnitudes (i.e., four different χ values) based on different Ψ values for the shaded cell. ${}^{F}\theta = 4$ and ${}^{H}\theta = 4$.

Figure 4.9 depicts the four qualitatively different ways in which a *high mismatch condition* can exist in a given CM_i. In panel a, both ${}_i\sigma$ and ${}_i\Omega$ are non-empty, however, they do not intersect. Thus, the cell implicated on the basis of the H-projection is different from that implicated by the F-projection. In this case, $\chi(x) = 0$ for all cells in the CM. Panel b depicts the case in which there is a strong Ψ value at one cell but no strong Φ values at any cell in the CM. Panel c depicts the complementary case in which a strong Φ value but no strong Ψ value exists. Finally, panel d shows the case where neither strong Φ nor strong Ψ values exist. Figure 4.10 is a more realistic depiction of each of these four high mismatch conditions. Note that in Figure 4.10, there are many non-zero χ values but assuming again that ${}^{H}\theta = 4$ and ${}^{F}\theta = 4$, they are all very small.

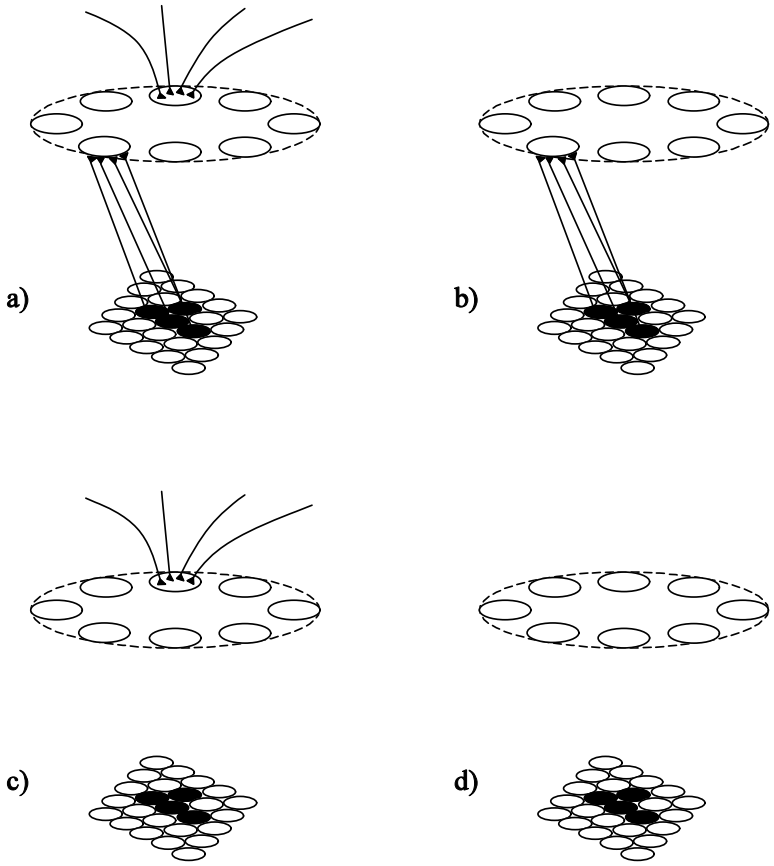

Figure 4.9: Four ways in which a high mismatch condition can exist in a given CM.

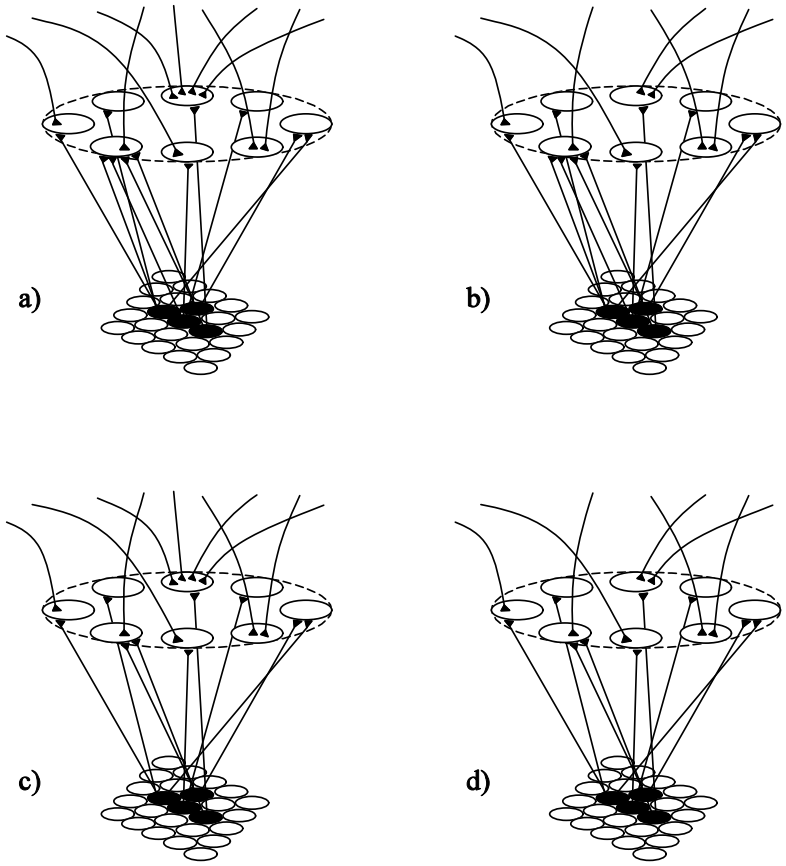

Figure 4.10: More realistic depiction of the four ways in which a high mismatch condition can exist in a given CM.

4.5.7 Compute normalized χ values

The χ values are then normalized in a fashion similar to the way in which the ϕ and ψ values are normalized. Let ${}_i\chi_t$ be the maximum χ value across all cells in CM_i on time slice *t*.

$$ {}_i\breve{\chi}_t = \max_{x \in CM_i} \chi_t(x) \tag{4.12}$$

The normalized χ values are:

$$ \mathrm{X}_t(x) = \frac{\chi_t(x)}{\max({}_i\breve{\chi}_t, {}^{\chi}\Theta_t)} \tag{4.13}$$

The reasoning for Eq. 4.13 is similar to that of the earlier analogous equations, Eq. 4.3 and 4.8. The threshold parameter, ${}^{\chi}\theta$, provides the system with another point of judgment—subsequent to the point at which prior context and current input information have been combined—as to the closeness of match between the current episode and previous episodes. If $\chi(x)$ is maximal in CM_i and $\chi(x) \geq {}^{\chi}\theta$, then x is considered to be perfectly matched to the current combination of prior context and current input information. The system treats this as maximal evidence that x should become active. Accordingly, $X_t(x) = 1.0$. ${}^{\chi}\theta$ interacts with the other threshold parameters in influencing the coarseness of the spatiotemporal categories formed by the model.

On the other hand, if no cell has $\chi \geq {}^{\chi}\theta$ then that indicates that no cell is maximally implicated by the current combination of prior context and current inputs. Accordingly, all x values for cells in this CM will be less than one. There will still, in general, be a distribution of x values reflecting the distribution of χ values, but the maximal x value will be less than one.

Note that having the highest x value in a CM still does not guarantee that a cell will be a final winner because of: a) the possibility of ties and, b) the possible addition of noise.

4.5.8 Compute the number of MCHs, Ξ_t, on the current time slice

Provided the maximal x value for CM_i is greater than or equal to ${}^{\chi}\theta$, all cells in CM_i tied for having the maximal x value in that CM are considered to represent multiple competing hypotheses—they are all equally implicated by the combined H- and F-signals. The system must decide how many MCHs exist on the current time slice in order to know how to set ${}^{H}\theta$ on the next time slice.

Let ξ_i be the number of cells, x, in CM_i, for which $\chi_t(x) = {}_i\breve{\chi}_t$. Then,

$$\Xi = round\left(\frac{\sum_{i=1}^{Q}\xi_i}{Q}\right) \tag{4.14}$$

where ‘round’ just means to round to the nearest integer.

4.5.9 Compute the final intra-CM degree of match, ${}_i\pi$

The final measure of match for CM_i is simply the maximum x value of the cells in CM_i.

$$ {}_i\pi_t = \max_{x \in CM_i} \mathrm{X}_t(x) \tag{4.15}$$

4.5.10 Compute overall degree of match, $\mathbf{G}_t$

The overall degree of match, G, at t is simply the average of the ${}_i\pi$'s.

$$ G = \frac{\sum_{i=1}^{Q} {}_i\pi_t}{Q} \tag{4.16}$$

G varies from 0.0, which indicates that the current input is completely unexpected based on the model's history of inputs (i.e., highly novel), to 1.0, which indicates that the current input is completely expected.

Although this match is appropriately described as being between the *expected* input and the *actual* input—i.e., as match in the space of L1 codes, L1-space, the actual matching takes place in the space of L2 codes, L2-space. The match is between normalized versions of the ϕ and ψ vectors.

4.5.11 Add noise into the winner selection process

Before defining the computation whereby an amount of randomness, dependent upon the overall degree of match, G_t, is added into the final choice of L2 winners at t, we reiterate the general goals of the winner selection process as explained in Sec. 4.1. They can be summarized as follows: The model's dynamics should achieve re-activation of old traces (i.e., pattern completion) in proportion to the familiarity of inputs and establishment of new traces in proportion to the novelty of inputs. In particular, as the similarity, $sim(\Gamma^j,\Gamma^i)$, between some novel episode, Γ^j, and the most-closely-

matching previously learned episode, Γ^i, increases, so does the overlap between Δ^j and Δ^i. Given that G measures $sim(\Gamma^j, \Gamma^i)$, the desired goal can also be stated as:

$$\left|\Delta_t^i \cap \Delta_t^j\right| \propto G \tag{4.17}$$

Computation of final probabilities of winning, $\hat{\rho}_t$

This section describes the computation whereby the desired relationship between G and the degree of overlap between L2 traces is achieved. The computation can be broken down into three stages.

a) Determine the range, $[\nu_{min}, \nu_{max}]$, into which the X values will be mapped.
b) Map the X values into ν values,
c) Map the ν values into the final probabilities of winning, i.e., the ρ values.

Stage one is defined as follows:

$$_R\nu = \frac{G^n \times \alpha \times K}{max(1, \Xi)} \tag{4.18}$$

where $_R\nu$ is the range of the ν distribution, $n \geq 1$ is an integer exponent governing the rate of increase of $_R\nu$ in G, α is a constant for linearly expanding the range of $_R\nu$, K is the number of cells per CM, and Ξ is the number of MCHs on the current time slice. ν_{min} is preset to a small, non-zero value—specifically, 1.0. ν_{max} is set to $\nu_{min} + {}_R\nu$.

In stage two, each of the x values are mapped through a sigmoid-shaped nonlinearity, depicted in Figure 4.11, to a corresponding ν value. The sigmoid function is defined as:

$$\nu_t(x) = \begin{cases} \nu_{\min} & X \le X_a \\ \nu_{\min} +_R \nu \times \dfrac{(X - X_a)^b}{(0.5 - X_a)^b + (X - X_a)^b} & X_a \le X \le 0.5 \\ \nu_{\min} +_R \nu \times \dfrac{(X_c - 0.5)^b}{(X_c - 0.5)^b + (X_c - X)^b} & 0.5 \le X \le X_c \\ \nu_{\min} +_R \nu & X > X_c \end{cases} \qquad (4.19)$$

where X_A is a threshold below which $\nu\_\min$ is returned, X_C is a threshold above which ν_{max} is returned, and b is an exponent governing the abruptness of the nonlinearity. As b increases, the function becomes closer and closer to a step function. L2 cells for which $X_t(x) \geq X_A$ will be assigned, in stage three, the smallest possible probability of winning. L2 cells for which $X_t(x) > X_C$ will be assigned the largest probability of winning. Thus, if more than one cell in a CM is tied for having the highest χ value in that CM (i.e., there are multiple competing hypotheses), then all such cells will be assigned equal probability of being chosen winner.

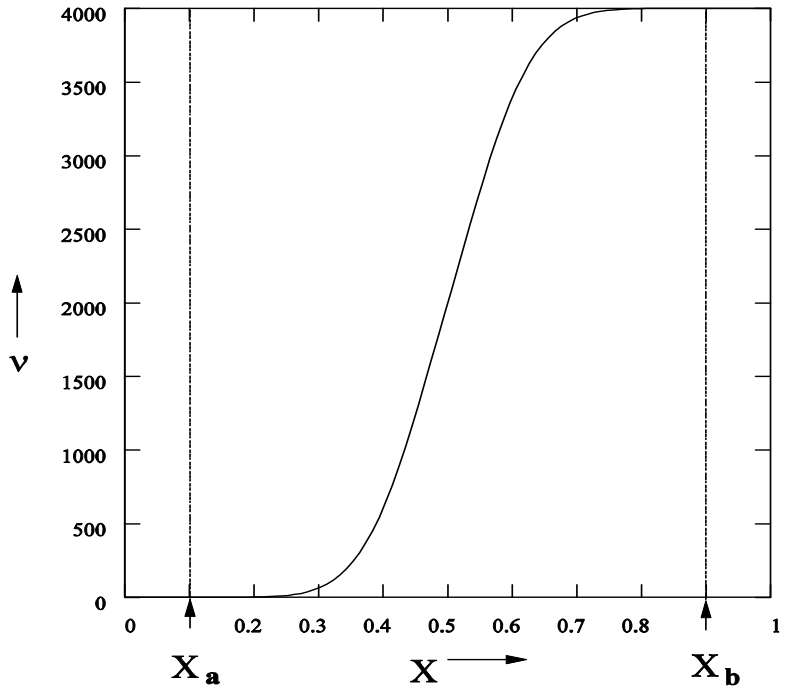

Figure 4.11: The sigmoid-shaped nonlinearity that maps x to ν.

In stage three, the final probabilities of winning—i.e., the ρ values—are computed as:

$$\rho_t(x) = \frac{\nu_t(x)}{\sum_{i=1}^{K} \nu_t(i)} \tag{4.20}$$

Together, the three stages of computation achieve the following. As G approaches one, ${}_R\nu$ is maximized. For example, the value of ${}_R\nu$ in Figure 4.11—i.e., 4000—was generated for the case of $G = 1.0$. The higher G is, the greater the disparity needed between ν_{max} and ν_{min}. On the other hand, as G approaches zero, ${}_R\nu$ approaches zero. In that limiting case, all x values map, via Eq. 4.19, to ν_{min}. This leads, in stage three, to an equal likelihood of winning for all cells in the CM under consideration. Thus, the overall goals of the winner selection process are met.

4.5.12 Compute final set of winners, Δ_t

The final step of the winner selection process is simply to choose actual winners according to the probability distribution, ρ. The finally chosen L2 code for t is denoted, Δ_t.

4.5.13 Compute learning rate parameter, η

In addition to all the previous steps, which are performed on each non-episode-initial time slice, while operating in the interactive mode, the model also computes the learning rate parameter, η, which depends inversely on G. In particular,

$$\eta_t = \begin{cases} 1.0 & G \le G_a \\ 1.0 - \dfrac{(G_c - 0.5)^b}{(G_c - 0.5)^b + (G_c - G)^b} & 0.5 \le G \le G_c \\ 1.0 - \dfrac{(G - G_a)^b}{(0.5 - G_a)^b + (G - G_a)^b} & G_a \le G < 0.5 \\ 0.0 & G < G_c \end{cases} \tag{4.21}$$

where G_A is a threshold below which 1.0 is returned, thus fully enabling learning at *t*, G_C is a threshold above which 0.0 is returned, thus fully disabling learning at *t*, and *b* is an exponent governing the abruptness of the nonlinearity. This direct control over the learning rate is needed to improve performance in handling sets of complex state sequences. All simulations reported in this chapter had parameter settings: $G_A = 0.2$, $G_C = 0.8$ and $b = 5$. Given that the model uses binary weights, the η parameter was used in the following probabilistic way. Any H-wt that is enabled for learning—i.e., that is currently equal to zero and whose presynaptic cell was active at *t-1* and whose postsynaptic cell is active at *t*—will be increased with probability, η_t.

4.6 Modified Algorithm for Solipsistic Mode

Sec. 4.5 described TEMECOR-II's interactive mode algorithm. That mode assumes that external input is present on every time slice of processing. This section briefly discusses the solipsistic mode algorithm. As discussed in Sec. 4.2, solipsistic mode is needed to explain the human information processing modes—i.e., thinking, reminiscence, imagination—in which external input is greatly attenuated and accordingly, during which the ongoing processing is primarily due to the propagating horizontal signals and not dependent on any matching between H- and F-expectancies.

In order to be meaningful, this mode must be entered by the system only after some L2 activity already exists. Otherwise, no L2 cells can become active while in this mode. Overall, the system is envisioned as operating in the interactive mode by default. However, it may enter the solipsistic mode at any time, t_{BS}, where '*BS*' stands for *begin solipsistic* phase. At some later time, t_{ES} (i.e., *end solipsistic* phase), the system begins accepting environmental input again and thus, re-enters interactive mode. At present, the theory does not contain a specific and comprehensive explanation of how the movement back and forth between interactive and solipsistic mode is controlled. In the simulations of Sec. 4.10.1, the model operates in the interactive mode while the prompt time slices are presented, and then enters the solipsistic mode during which period it attempts to read out the remainder of the trace for the prompted episode.

Two definitions are necessary in order to describe the desired properties of solipsistic mode. Let us define the concept of a *familiar L2 code* as a non-initial instance of that L2 code. That is, an

instance of an L2 code is familiar if there exist previous instances of that L2 code having been active during the *life* of the system. More generally, a particular instance of a sequence of L2 codes is familiar if it is a non-initial instance of that sequence. Let us define a *novel L2 code* (or sequence of L2 codes) as one that is not familiar. Given these definitions, the primary desired properties of solipsistic mode are:

a) If the L2 code, Δ_t^*, active immediately prior to the beginning of a solipsistic phase was familiar and if no noise is added during the solipsistic phase, then the sequence of L2 codes that occurs during the solipsistic phase should be identical to the remainder of Δ_t; i.e., Δ_{t+1}^* through the final time slice of that episode, Δ_f^*. This corresponds to *reminiscence mode*.

b) Otherwise, if either the L2 code active immediately prior to the beginning of a solipsistic phase was not familiar (i.e., was novel) or if noise *is* added during the solipsistic phase, then the sequence of L2 codes that obtains during the solipsistic phase will have some, perhaps large, degree of novelty. This case affords the opportunity for new learning despite the fact that there is no input to the system during the period. This corresponds to the *fantasizing* described briefly in Sec. 4.2. This mode is not explored in this thesis.

The preceding list of desired properties implies that noise may be present during a solipsistic phase. Yet, we have described the amount of noise added to the winner selection process as being dependent upon the degree of match between the H-expectancy and the F-expectancy. However, in solipsistic mode, there is no F-expectancy. Thus, another source of control over the level of noise would have to be assumed for solipsistic mode. This is a matter for future research.

The changes to the interactive mode's processing algorithm are as follows. Since there is no F-input, steps one and two are not performed. Stages six through fifteen are replaced by a single step in which the cell in each CM receiving the most H-input is simply chosen as winner.

$$\Delta_t(i) = \left\{ x \mid \phi_{t(x)} =_i \breve{\phi}_t \right\} \tag{4.22}$$

Ties are broken at random. Note that step five (Φ vector) is also not necessary since the cell with maximal ϕ is also the cell with maximal Φ. Also note that noise is not modeled in solipsistic mode.

Finally, the following two steps are needed to explain how the correct L1 code can be reinstated from the active L2 code. On each time slice, the total R-input to each L1 cell, from the active L2 cells, is computed as:

$$\zeta_{t(x)} = \sum_{j \in \Delta_t} w_{jx} \tag{4.23}$$

Then, all L1 cells having a ς value that meets or exceeds are activated on this time slice.

$$\Gamma_t = \left\{ x \mid \zeta_t(x) \geq^{R} \theta \right\} \tag{4.24}$$

4.7 Algorithm Summary

4.7.1 Interactive mode

1. $\psi_{i,t} = \sum_{j \in \Gamma_t} w_{ji}$ (4.1)
2. $\Psi_t(x) = \dfrac{\psi_t(x)}{\max(_i \breve{\psi}_t, {}^{F}\Theta_t)}$ (4.3)
3. If $t > 0$: ${}^{H}\theta_t = \dfrac{{}^{H}\theta_{baseline}}{\Xi_{t-1}}$ (4.5)
4. If $t > 0$: $\phi_t(x) = \sum_{j \in \Delta_{t-1}} w_{jx}$ (4.6)
5. If $t > 0$: $\Phi_t(x) = \dfrac{\phi_t(x)}{\max(_i \breve{\phi}_t, {}^{H}\Theta_t)}$ (4.8)
6. $\chi_{i,t} = \begin{cases} \Psi_{i,t}{}^{u} \Phi_{i,t}{}^{v} & ,t > 0 \\ \Psi_{i,t}{}^{w} & ,t = 0 \end{cases}$ (4.11) (4.10)
7. $\mathrm{X}_t(x) = \dfrac{\chi_t(x)}{\max(_i \breve{\chi}_t, {}^{\chi}\Theta_t)}$ (4.13)

8. $$\Xi = round\left(\frac{\sum_{i=1}^{Q}\xi_i}{Q}\right) \tag{4.14}$$

9. $$_i\pi_t = \max_{x\in CM_i} \mathrm{X}_t(x) \tag{4.15}$$

10. $$G = \frac{\sum_{i=1}^{Q} {}_i\pi_t}{Q} \tag{4.16}$$

11. $$_R\nu = \frac{G^n \times \alpha \times K}{max(1,\Xi)} \tag{4.18}$$

12. $$\nu_t(x) = \begin{cases} \nu_{\min} & \mathrm{X} \le \mathrm{X}_a \\ \nu_{\min} +_R \nu \times \dfrac{(\mathrm{X}-\mathrm{X}_a)^b}{(0.5-\mathrm{X}_a)^b + (\mathrm{X}-\mathrm{X}_a)^b} & \mathrm{X}_a \le \mathrm{X} \le 0.5 \\ \nu_{\min} +_R \nu \times \dfrac{(\mathrm{X}_c-0.5)^b}{(\mathrm{X}_c-0.5)^b + (\mathrm{X}_c-\mathrm{X})^b} & 0.5 \le \mathrm{X} \le \mathrm{X}_c \\ \nu_{\min} +_R \nu & \mathrm{X} > \mathrm{X}_c \end{cases} \tag{4.19}$$

13. $$\rho_t(x) = \frac{\nu_t(x)}{\sum_{i=1}^{K}\nu_t(i)} \tag{4.20}$$

14. Choose Δ_t according to the ρ distribution.

15. $$\eta_t = \begin{cases} 1.0 & G \le G_a \\ 1.0 - \dfrac{(G_c-0.5)^b}{(G_c-0.5)^b + (G_c-G)^b} & 0.5 \le G \le G_c \\ 1.0 - \dfrac{(G-G_a)^b}{(0.5-G_a)^b + (G-G_a)^b} & G_a \le G < 0.5 \\ 0.0 & G < G_c \end{cases} \tag{4.21}$$

4.7.2 Solipsistic mode: non-episode-initial time slices

The solipsistic mode's entire processing algorithm consists of five steps. η does not need to be set because we are not considering the case in which learning is allowed in solipsistic mode.

1. $${}^{H}\theta_t = \frac{{}^{H}\theta_{baseline}}{\Xi_{t-1}}$$ (4.5)

2. $$\phi_t(x) = \sum_{j \in \Delta_{t-1}} w_{jx}$$ (4.6)

3. $$\Delta_t(i) = \left\{ x \mid \phi_{t(x)} =_i \breve{\phi}_t \right\}$$ (4.22)

4. $$\zeta_{t(x)} = \sum_{j \in \Delta_t} w_{jx}$$ (4.23)

5. $$\Gamma_t = \left\{ x \mid \zeta_t(x) \geq {}^{R}\theta \right\}$$ (4.24)

4.8 Traces of TEMECOR-II algorithm

This section contains traces of several scenarios that reveal the various properties listed in Sec. 4.3. The reader can assume that ${}^{H}\theta = Q\text{-}1 = 5$, ${}^{F}\theta = S = 4$, ${}^{\chi}\theta = 0.85$, $u = 2$, $v = 2$, $w = 2$, and $b = 2$ throughout these examples, unless otherwise stated.

4.8.1 Example 1: Presentation of single state, A

This example describes the learning that takes place, in the F-projection, between an L1 code, Γ_A (state A), and its associated L2 code, Δ_A. This example can be thought of as a degenerate episode consisting of a single time slice that is therefore also an episode-initial time slice. Figure 4.12 shows how the L1 code, Γ_A, having $S = 4$ active cells, gets linked to a particular L2 code, Δ_A. The model in this and subsequent examples has $Q = 6$ CMs. Assuming this is the first time slice of this model's existence, all H- and F-weights are zero. Thus the ψ and ϕ values for all L2 cells are zero. Thus, by Eq. 4.10, χ values are all zero. Thus, ${}_i\pi = 0$, $\forall \mathrm{CM}_i$. Therefore $G_A = 0$. This implies that the set of winners (black cells of L2) depicted in the figure is purely the result of noise. Hebbian learning then occurs between the L1 and L2 codes; both in the F- and R-weights. Note that the solid vertical lines in this and subsequent similar figures represent the F-weights and the corresponding R-weights.

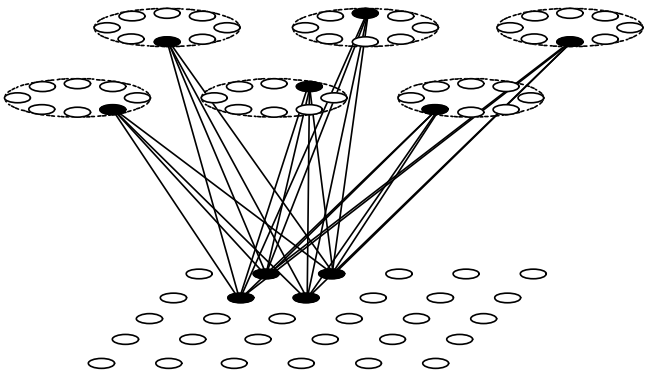

Figure 4.12: The linkage between an L1 code, Γ_A, and an L2 code, Δ_A. The lines denote bidirectional connections, both of which would be increased to a weight of 1 in this example. A total of 24 (i.e., 4 L1 cells $\times$ 6 L2 cells) F-connections and the 24 corresponding R-connections are increased in this case.

We now describe how future presentations of state A and of degraded versions of state A, in an episode-initial position, will interact with L2.

Subsequent episode-initial occurrences of Γ_A would result in $\psi = S = 4$ for the six black L2 cells—i.e., Δ_A—in Figure 4.12. Assuming ${}^F\theta$ = S, then by Eqs. 4.3-4.16, there would be a very high degree of match—in fact, $G = 1.0$ and no noise would be added into the winner selection process. Therefore, the winning L2 code would be determined purely on the basis of the deterministic, experience-dependent F-signals. Accordingly, Δ_A would be reinstated. The lack of noise in the winner selection process in this instance leads to the desirable result that no new L2 cells are used to represent this familiar input. This example shows that TEMECOR-II has the property episode-initial L1 codes can correctly elicit the correct episode-initial L2 codes.

Interference from other previously learned vertical mappings

The previous example is a particularly ‘clean’ example in that there are no other previously learned L1-to-L2 associations that could potentially interfere with reinstatement of the correct L2 code. Suppose the L1-to-L2 association—say, from an L1 state, Γ_Y, to an L2 code, Δ_Y—shown in Figure 4.13 had been learned in the past. Note in particular that Γ_A has one cell in common with Γ_Y. Note also that $|\Delta_A \cap \Delta_Y| = 1$. The reader can assume that the overlap of the L2 codes is due to chance. Figure 4.14 shows the vector of F-inputs, \hat$\psi$, when Γ_A is reinstated. Although there are

non-zero F signals arriving at the L2 cells (light gray) that are members of Δ_Y (i.e., spurious signals), they only have χ values of $(0.25)^2 = 0.0625$ (by Eq. 4.10), whereas the correct L2 cells (i.e., those in Δ_A) have $\chi = (1.0)^2 = 1$. Following passage through the subsequent non-linearity, Eq. 4.19, the computation of the final probabilities of winning, ρ, in Eq. 4.20, will overwhelmingly favor the cells of Δ_A (black cells). Thus we expect reinstatement of Δ_A with very high likelihood despite the interference from the other learned mapping.

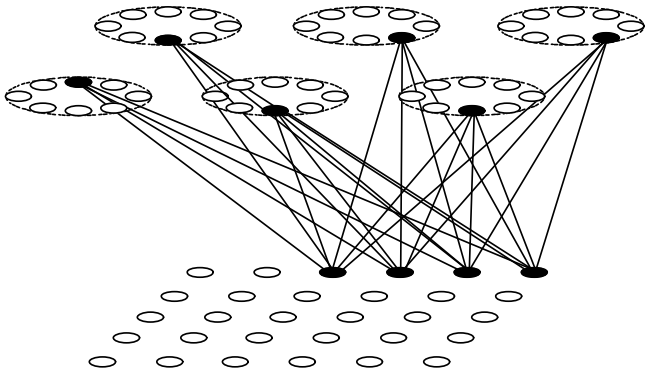

Figure 4.13: Another vertical mapping. This one is between an L1 state, Γ_Y, and its L2 code, Δ_Y. Because $|\Gamma_A \cap \Gamma_Y|$ is non-empty, there will be spurious F-signals when either L1 state is reinstated, however, as seen in the next figure, the correct F-signals will dominate and the correct L2 code will become active.

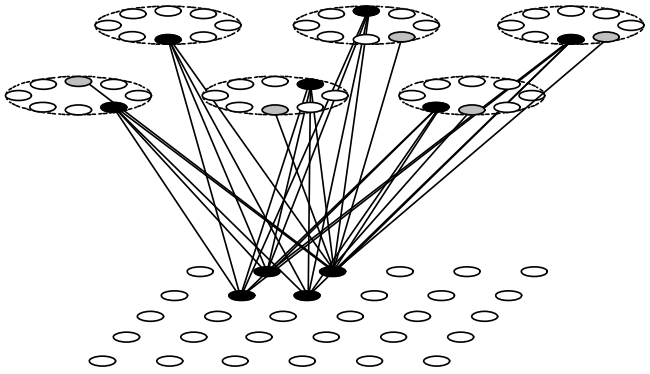

Figure 4.14: When state A is reinstated at L1, there are some spurious signals generated at L2, due to interference from the mapping between Γ_Y and Δ_Y, however, the L2 cells in Δ_A (black cells) have much larger inputs than those of Δ_Y (gray cells) and will, with very high probability, become winners in their respective CMs.

Degraded (Partial) prompts: spatial pattern completion

Suppose that after having learned the vertical—i.e., spatial—association, Γ_A to Δ_A, we subsequently reinstate a degraded version of state A that has missing features. Specifically, suppose a state, A′, having three cells in common with state A, is presented to the model, as shown Figure 4.15.

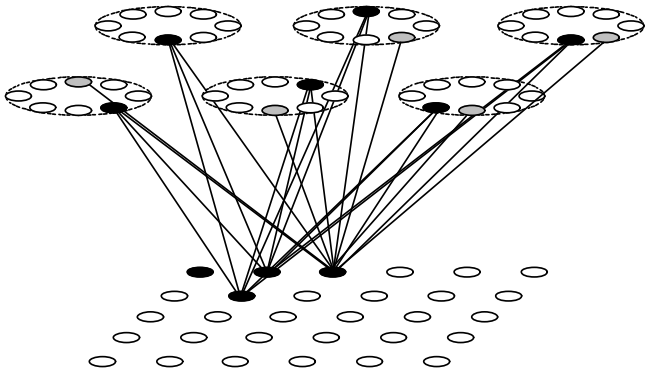

Figure 4.15: Reinstatement of a degraded (partial) version, A′, of state A which is missing one of state A's four features leads to smaller χ values (i.e., a worse match) and thus the presence of some noise. The vector of F signals, $\hat{\psi}$, clearly favors the correct L2 code—i.e., all the correct L2 cells have $\psi = 3$ whereas the incorrect cells have $\psi = 1$. The amount of noise added depends on various model parameters. For example, if ${}^{F}\theta$ were set to 3, rather than S = 4, no noise would be added. The exponent parameter, w, for the nonlinearity, Eq. 4.10, also influences the final amount of noise added into the winner selection process. For the parameters we have assumed so far, the cells in Δ_A would each have $\chi = (3/4)^2 = 0.5625$ and those of Δ_Y would have $\chi = 0.0625$. Subsequent steps of the algorithm, operating on these values would lead to reinstatement of Δ_A with fairly high probability.

In this case, $\chi = (3/4)^2 = 0.5625$ for the cells of Δ_A (black cells) and $\chi = (1/4)^2 = 0.0625$ for the cells of Δ_Y. This leads to *x* values of 0.6617 and 0.0735, respectively and thus to $G = 0.6617$. This has the effect, in Eq. 4.18 of reducing the range of ν values to which the *x* values will be mapped. In particular, assuming $\alpha = 100$ and $n = 2$, ${}_R\nu \approx 350$. This reduces the range of final probabilities of

winning of the cells in any given CM. The maximal possible value for ${}_R\nu$, when $G = 1.0$, is 800. After processing through Eqs. 4.19 and 4.20, the final probabilities, are, 96.8% for the cells of Δ_A, 0.7% for the cells of Δ_Y, and 0.4% for the other, as-yet-unused L2 cells of the figure. Thus, for these particular parameters and history of inputs, Δ_A is very likely to be perfectly reinstated.

This example shows that the model is capable of performing pattern completion in the spatial domain. If the degraded input, A′, does actually lead to Δ_A, then the model is effectively co-categorizing the new input with the old.

Continuity in the L1-to-L2 mapping

Now suppose that an even more degraded version, A″, of state A is re-presented to the network. As shown in Figure 4.16, A″ has only two cells in common with A. In this case, $\chi = (2/4)^2 = 0.25$ for the cells of Δ_A and, as in the previous example, $\chi = (1/4)^2 = 0.0625$ for the cells of Δ_Y. This leads to x values of 0.294 and 0.0735, respectively, and thus to G = 0.294. This leads to ${}_R\nu \approx 69$. The final probabilities of winning are 72.2% for the cells of Δ_A, 4.4% for the cells of Δ_Y, and 3.9% for the six other, as-yet-unused L2 cells of the figure. Thus, the lesser degree of match in this example relative to the previous example leads to a greater expected degree of deviation between the final set of winners chosen and the set of most-strongly-implicated cells—i.e., Δ_A. Recall from Sec. 4.1 that we refer to an instance in which the most-strongly-implicated cell, on the basis of deterministic influences, is not the final winner as an instance of *winner flip*. For the ρ values in this example, we expect winner flip in about three out of every ten CMs. Figure 4.16 shows the case in which winner flip has, in fact, occurred in two of the six CMs. The copy of L2 that appears above L2 denotes the finally chosen L2 code (black cells). The light gray L2 cells denote members of Δ_Y and have the least support. The dark gray cells denote members of Δ_A and have more support. The asterices denote the two instances in which the purely deterministic choice of winner (i.e., based on $\hat{\psi}$) has been overridden by the noise.

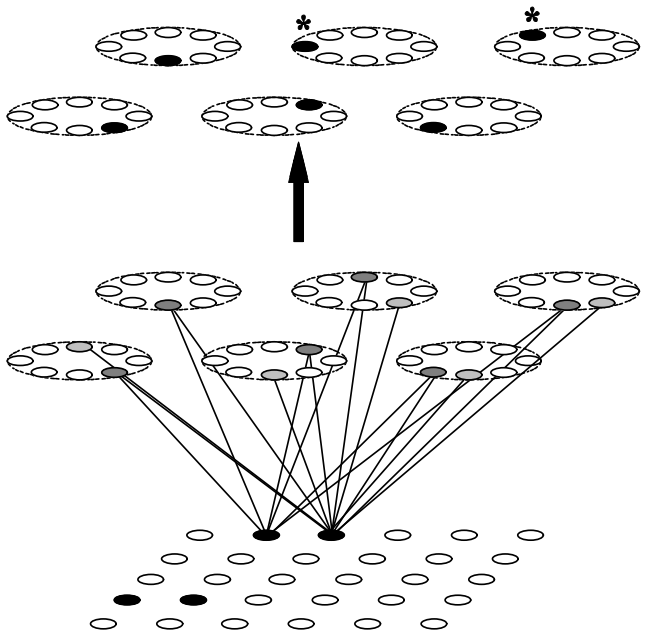

Figure 4.16: Reinstatement of a more degraded version, A′, of state A having only two of four features in common with state A. All π values equal 0.294, as does G, and more noise is added than in the previous figure. $\hat{\psi}$ still favors the correct L2 code however, due to greater noise, the final winner is different from the purely deterministically chosen winner in two of the six CMs. Light gray cells are members of Δ_Y. Dark gray cells are members of Δ_A. The copy of L2 that appears above L2 denotes the finally chosen L2 code (black cells). The asterices denote the two instances in which the final winner is not the one implicated by $\hat{\psi}$.

Figure 4.17 shows reinstatement of a still more degraded version, A′′′, of state A, having only one feature in common with A. In this case, the match is indeed very poor and the probabilities of winning for the most-strongly-implicated cells, which is still the set, Δ_A, are only slightly larger than the probabilities of winning for the other cells. Thus we expect the overlap between the finally chosen L2 code and Δ_A to be much closer to that corresponding to chance. Accordingly, Figure 4.17 depicts a final L2 code (black cells) having only one cell in common with Δ_A (gray cells).

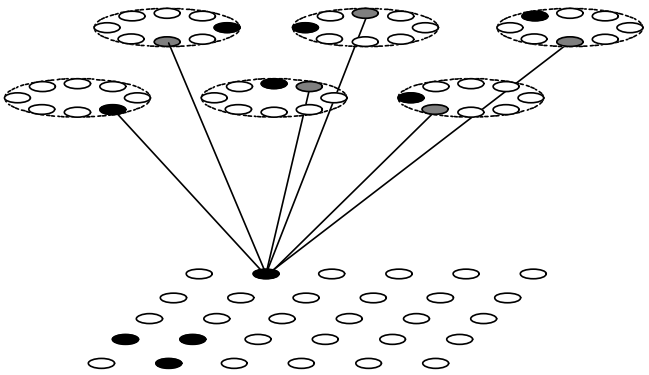

Figure 4.17: Reinstatement of an even more degraded version, A‴, of state A having only one feature in common with state A. The degree of match is smaller still than that of the previous example and the finally chosen L2 code has only one cell (in the lower leftmost CM) in common with Δ_A. Winner flip has occurred in the other five CMs. Dark gray cells are members of Δ_A. Black cells are final winners (including the influence of noise).

This extended example involving degraded versions of state A has shown that as we consider states having progressively less overlap with state A, progressively more noise is added into the winner selection process, resulting in progressively less similarity between the finally chosen L2 code and Δ_A. Thus, this example demonstrates that the relationship, 4.17, holds for the purely spatial domain. The simulation results of Sec. 4.10.2, in which we show that amount of learning varies inversely with *G*, provides evidence that this proportionality holds for the more general spatiotemporal case as well.

4.8.2 Example 2: Presentation of simple state sequence, [A B]

The previous example involved only the F-projection. This example involves both the F- and H-projections and their interrelationship. In particular, this example describes the formation of an L2 memory trace in response to presentation of the simple state sequence, Γ^1 = [A B].

The discussion of the choice of an L2 code for state A is qualitatively similar to that given at the beginning of Sec. 4.8.1 and will not be repeated here. Thus, Figure 4.12 accurately depicts the L2 code, Δ_A, which will now be more precisely denoted as Δ^1_A (or Δ^1_1) that gets chosen on the first time slice of the current example. (Note, however that this example does not assume that Γ_Y has occurred previously.)

Figure 4.18 shows the L1 code corresponding to state B having two L1 cells (features) in common with state A. Since $\left|\Delta_A^1 \cap \Delta_B^1\right| = 2$, each of the L2 cells that were members of Δ_A^1 will have two active learned inputs via the F-projection—i.e., $\psi = 2$. These cells are depicted in light gray to indicate that they are merely receiving some support but will not ultimately become active on this time slice.

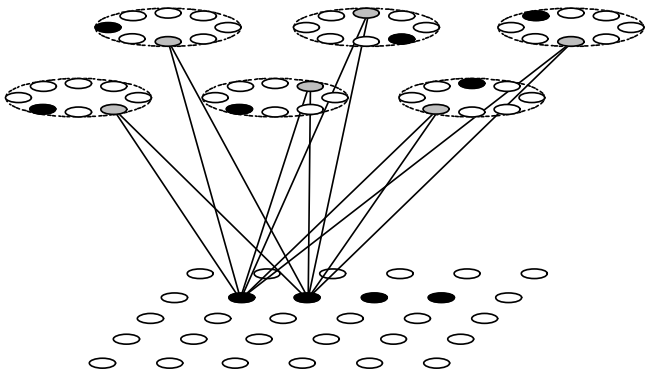

Figure 4.18: A state, B, having an overlap of two L1 cells with state A is shown. This overlap means that each of the cells in the L2 code previously chosen for state A (light gray) will have a ψ value of 2. Assuming that this is the first episode presented to the model, all H-wts are zero. This means that the ϕ and Φ values for all cells are also zero. This further implies, according to Eq. 4.11 that all χ values are 0. Thus G = 0 and all L2 cells are equally likely to be chosen winner to represent state B. That is, noise completely predominates in the winner selection process. Thus, the finally chosen L2 code for state B, Δ_B^1 (black cells), has zero overlap with Δ_A^1.

Since, by assumption, this is only the second time slice ever experienced by the model, no horizontal weights can have been increased yet. Thus, the ϕ and Φ values for all cells are also zero. This further implies, according to Eq. 4.11 that all χ values are 0. Thus $G = 0$. By Eq. 4.18, ${}_R\nu = 0$. Thus, by Eqs. 4.19 and 4.20, all L2 cells are equally likely to be chosen winner to represent state B. $G = 0.0$ has led to the winner choice process being completely dominated by noise. The resulting L2 code for state B, Δ_B^1 (black cells), will be highly unique. In this case in particular, we have chosen a hypothetical Δ_B^1 having zero overlap with Δ_A^1.

Following the choice of $\Delta_B^1 = \Delta_2^1$, hebbian learning will take place in both the F- and H-projections. A few of the horizontal synapses from cells in Δ_A^1 onto Δ_B^1, which would be increased, are shown Figure 4.19. This figure also shows some of the F-weights (and thus, R-weights) that would be increased.

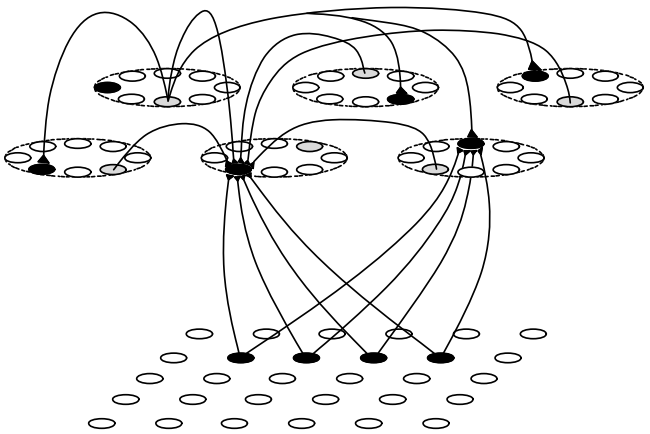

Figure 4.19: A new state, B, is active at L1. The previous L2 code, Δ_A^1, is depicted in gray. The newly chosen L2 code, Δ_B^1, is represented by the black cells of L2. A suggestive subset of those H- and F-weights that would be increased on this second time slice of I^1 is shown. Note that actual synapses are not shown in this figure.

4.8.3 Example 3: Re-presentation of familiar episode, [A B]

Now suppose the next episode presented to the model, Γ^2, is identical to Γ^1 (i.e., [A B]). In this case, when Γ_1^2 = A presents, all of the L2 cells that were members of Δ_1^1 will have $\psi = 4$ due to the learning in the F-projection that occurred during the first presentation of A under Γ^1 (shown in Figure 4.12). Therefore, by Eq. 4.10, the χ values for all the cells in Δ_1^1 will equal 1.0. This ultimately leads to extremely high probabilities of winning for the cells in Δ_1^1 and near-zero probabilities of winning for the other cells. In other words, nearly zero noise is added into the winner selection process in this case and the probability of winner flip is also nearly zero. The resulting L2 code, Δ_1^2, will in all likelihood be identical to Δ_1^1.

On the next time slice, when Γ_2^2 = B presents, the L2 code, Δ_2^1 (i.e., the black L2 cells of Figure 4.19), will be strongly implicated by the vector of H-inputs from the L2 cells active on the previous

time slice (i.e., Δ_1^1)—in fact, each of the cells that were members of Δ_2^1 will have $\phi = Q\text{-}1 = 5$ and $\Phi = 1.0$. Furthermore, this same set of L2 cells will be strongly implicated by the vector of F-inputs (due to the learning in the F-projection that would have obtained while processing the initial instance of this episode). Thus, in all CMs, there will be a very strong match condition of the type shown in Figure 4.6b. Specifically, the χ's for all cells in Δ_2^1 will equal 1.0 and therefore so will G_2^2. This results in reactivation of Δ_2^1 with probability very close to 1.

This is an example in which the model is functioning in the *interactive tracking mode* discussed earlier. No new learning in either the H- or F-projections obtains during processing of Γ^2. This is because the succession of L2 codes—i.e., *L2 swath*—that arises during processing of Γ^2 is identical to a swath that has occurred at some point in the past; in particular during presentation of Γ^1. Because the model assumes that synapses increase to asymptote of 1.0 on the first occasion on which they increase at all, exact reactivation of such a *familiar* L2 swath affords no opportunity for new synaptic increases. More generally, the degree of overlap between the L2 swath that is currently obtaining and the set of L2 swaths that have occurred in the past governs the amount of learning that can take place in the current instance. Example one showed that the degree of overlap between L2 codes is an increasing function of the similarity (degree of overlap) of the corresponding L1 codes. Putting these two facts together shows that the model exhibits the psychologically plausible property that the amount of learning is an increasing function of the degree of novelty. This concept is discussed again in Sec. 4.9.

Moreover, no external signal telling TEMECOR-II whether the current trial is a learning or a recall trial is necessary in this example. By controlling the degree to which old traces are reactivated, the instantaneous global degree of match, *G*—which is an internally generated signal—places an upper bound on the degree of learning possible on any given time slice.

4.8.4 Example 4: A complex sequence set, [A B C] and [D B E]

Now consider how the model would handle the two sequences,

Γ^1: [A B C]
Γ^2: [D B E]

that together constitute a complex sequence set. The explanation of the processing of Γ^1 would be similar to previous explanations in examples 1–3. Note that this example assumes a new (i.e., unused) network. Assuming the three states, A, B and C are fairly different from each other, it follows that the ϕ signals present on time slices, *t = 2* and *t = 3*, will be very weak, thus that χ values, even for the most-strongly-implicated L2 cells, will be very low, thus the winner selection process on these time slices will be dominated by noise, and finally, that highly distinct L2 codes will be chosen for each of these input states. In particular, we assume that the L2 codes chosen for states A and B are the same as those depicted in Figures 4.12 and 4.18. Δ^1_C can be assumed to be some other L2 code having very low overlap with any other L2 code. Δ^1_C is not depicted in any figure.

Then Γ^2 is presented. We assume that states D and E are also fairly different from each other and from A, B, and C. Therefore it follows that the ψ values when D presents will be small, and that G_1^2 will be low, and thus that $\left|\Delta^1_A \cap \Delta^2_D\right|$ will be small. For the sake of example, we assume that $\left|\Delta^1_A \cap \Delta^2_D\right| = 0$. Recalling our assumption that ${}^H\theta = Q\text{-}1$, this null intersection implies that when state B presents at t = 2, all ϕ (and Φ) values will be zero. On the other hand, the vector of F-inputs at t = 2 will strongly implicate the L2 code learned for state B during Γ^1—i.e., Δ^1_B. In fact, all of the cells in Δ^1_B will have ψ values of four and Ψ values of 1.0. Nevertheless, because of the zero H-inputs, all χ values will be zero (by Eq. 4.11), *G* will equal zero, the winner selection process will be completely dominated by noise, and the resulting L2 code, Δ^2_B, will have chance-level intersection with Δ^1_B.

Learning in the H-projections—between the cells comprising Δ^2_D and Δ^2_B—will then take place. Learning in the F-projection—between the cells of Γ^2_B and Δ^2_B—also takes place at this time. Figure 4.20 shows hypothetical choices of L2 codes for the first two states of the sequence [D B E] and some of the associated learning in the F- and H-projections. Note that $\left|\Delta^1_A \cap \Delta^2_D\right| = 0$ and $\left|\Delta^1_B \cap \Delta^2_B\right| = 0$.

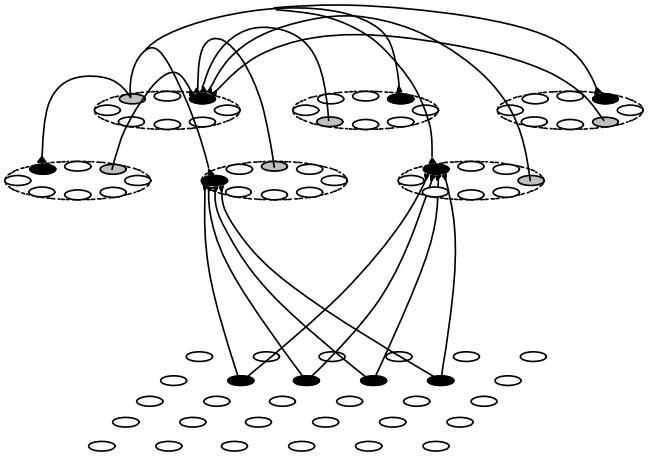

Figure 4.20: This figure shows a possible choice of L2 codes for the first two states of the sequence [D B E]. The light gray cells represent $\Delta_D^2 = \Delta_1^2$ ***and the black cells represent*** $\Delta_B^2 = \Delta_2^2$***. Some of learning in the F- and H-projections that would take place at the end of t = 2 is shown.***

Assuming that many of the cells chosen for Δ_B^2 are active for the first time in Δ_B^2, they will have had no opportunity to increase any of their outgoing H-wts. Thus the ϕ values at *t = 3* will probably be approximately zero. This implies that $G_3^2 \approx 0$, and ultimately that another highly distinctive L2 code, Δ_E^2, will be chosen for state E. Learning between the cells comprising Δ_B^2 and Δ_E^2 will then take place at *t = 3*.

The main point to notice in this example is that the internal representation (L2 code) for state B is very different in each episode and that this difference resulted from the different prior contexts, A vs. D, in the two cases. This example parallels the Ch. 3 example describing how TEMECOR-I is able to learn two sequences having a common state.

Subsequent re-presentation of state A will lead to reinstatement of the L2 code, Δ_A^1. Δ_A^1 will, as a consequence of the learning in the H-projection that took place during processing of Γ^1, give rise to a vector of H-inputs, $\hat{\phi}$, at *t = 2*, that strongly implicates Δ_B^1. At this point, it becomes necessary to consider three possible scenarios:

a) The model remains in interactive mode and state B presents at *t = 2* just as it did when the Γ^1 initially occurred.

b) The model remains in interactive mode and a novel state, F, different from any of the other states, A–E, occurs.
c) The model enters solipsistic mode (disregarding input). In this case, we'd like the model to still read out the remainder of Γ^1, i.e., [B C].

Interactive mode: state B presents

If state B presents following A, the same set of cells, Δ^1_B, will be maximally implicated by the combined H- and F-inputs. The match situation will be as depicted in Figure 4.7c. *G* will equal 1, near-zero noise will be added into the winner selection process, and with very high probability, Δ^1_B will be fully reinstated. If C then presents on the next time slice, the H- and F-expectancies will again match very strongly and Δ^1_C will be reinstated. As in example 2, no learning will occur during processing of this repeat occurrence of Γ^1. This would also be an instance of *interactive tracking* mode.

Similarly, if the model had been prompted with state D, the first time slice of Γ^2, then Δ^2_D would have become active at *t = 1*. In this case, an H-expectancy identical to Δ^2_B (rather than to Δ^1_B) would have been present at the beginning of *t = 2*. In this case, when state B presents, then Δ^2_B will become active. Finally, if state E follows at *t = 3*, Δ^2_E will become reinstated. Thus, both sequences, [A B C] and [D B E], can be recalled without interfering with each other.

Interactive mode: a novel state occurs

On the other hand, suppose a novel state, F, that differs substantially from all states A–E, occurs following A. In this case, the H- and F-expectancies will substantially mismatch. This situation is depicted in Figure 4.21. The depicted L1 code, corresponding to F, Γ^F, has one cell (the rearmost black cell) in common with state B (see Figure 4.18). That cell is also in common with state A (see Figure 4.12). Assume the other three cells in Γ^F have never been active before.

The dark gray cells in the figure denote the set, Δ^1_B. These are the most-strongly-implicated cells at this point in time. The light gray cells denote Δ^1_A. They are receiving some minimal support

via the F-projection. Note also that the light gray cells were active on the previous time slice. Even the Δ_B^1 cells, which have high ϕ values (indicated by the five large active H-synapses impinging on each of the cells), have only small χ values (0.0625) according to Eq. 4.11). Thus, the overall degree of match, *G*, also equals 0.0625. This leads to final probabilities of winning that only slightly favor the cells of Δ_B^1 over those of Δ_A^1 or even over the as-yet-unused cells. That is, much noise will be present in the winner selection process and the probability of winner flip will be very high. Accordingly, the finally chosen L2 code (black L2 cells) in Figure 4.21 overlaps with Δ_B^1 at only one cell and overlaps with Δ_A^1 at no cells.

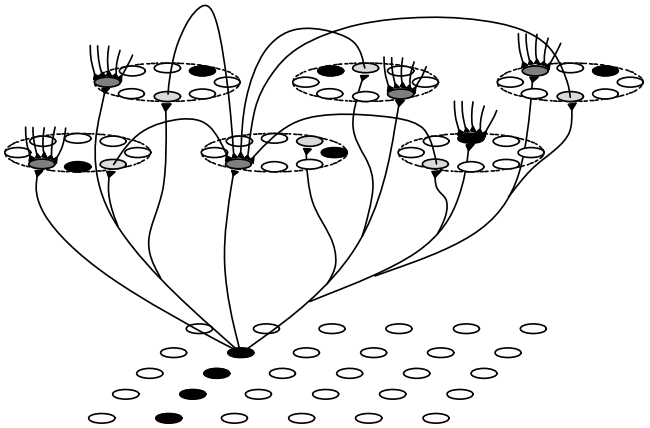

Figure 4.21: Strong mismatch despite strong implication by previous L2 code. The L1 code represents a novel state, Γ_F. The rearmost active (i.e., black) L1 cell was also contained in two prior L1 codes, Γ_A and Γ_B, and had its F-wts onto Δ_A^1 (light gray L2 cells) and Δ_B^1 (dark gray L2 cells) increased accordingly. The cells of Δ_B^1 are highly favored on the basis of H-inputs but no cells are strongly favored on the basis of F-inputs. Since the final χ values reflect a multiplicative interaction between the F- and H-inputs, no cells end up indicating a strong match condition. Accordingly, G is near zero, noise dominates the winner selection process and the finally chosen L2 code has low intersection with any preexisting L2 codes. Note that the full axons reaching all the way from a light gray cell to a dark gray cell are shown only for one of the dark gray cells. Only partial axons are shown for each of the other dark gray cells in order to keep the figure readable.

Model enters solipsistic mode

Alternatively, if the input were suddenly to be turned off following presentation of state A, the model could *enter solipsistic recall* (i.e., *reminiscence*) mode. In this case, the model has enough information to uniquely determine the rest of the sequence to read out and it should be capable of doing this. No matching between F- and H-signals occurs in solipsistic mode. In addition, no noise is added into the winner selection process in this mode. Rather, successive L2 winners are simply those whose ϕ values a) are maximal within their respective CMs, and b) meet or exceed ${}^{H}\theta$. As explained in the immediately previous subsection, $\phi = 5 = {}^{H}\theta$ for all L2 cells that were members of Δ^1_B. Thus, Δ^1_B will be reactivated. By similar reasoning, Δ^1_C will become active at $t = 3$ if the model remains in solipsistic mode.

In addition to explaining how successive L2 codes can be correctly read out in solipsistic mode, we must also explain how the corresponding L1 codes can also be reinstated. This is accomplished by having a threshold parameter for the inputs to the L1 cells via the reciprocal, R-projection. Specifically, on each time slice, all L1 cells for which the total summed R-input meets or exceeds ${}^{R}\theta$ are activated. Thus, during solipsistic recall, each L2 code elicits a) the subsequent L2 code, if there was one, and b) the contemporaneous L1 code.

4.8.5 Example 5: Ambiguous prompt

This example describes how an ambiguous prompt causes multiple competing hypotheses (MCHs) to become active simultaneously and how the subsequent presentation of disambiguating information resolves this competition leaving only the consistent hypothesis active. Suppose we present the prompt, B, which is ambiguous given the two episodes experienced in the last example, [A B C] and [D B E]. As shown in Figure 4.22, the cells in both Δ^1_B and Δ^2_B are equally strongly implicated by the F-inputs. Since Δ^1_B and Δ^2_B are disjoint (this can be checked by comparing the L2 codes in Figures 4.18 and 4.20, there are two dark gray cells in each CM.

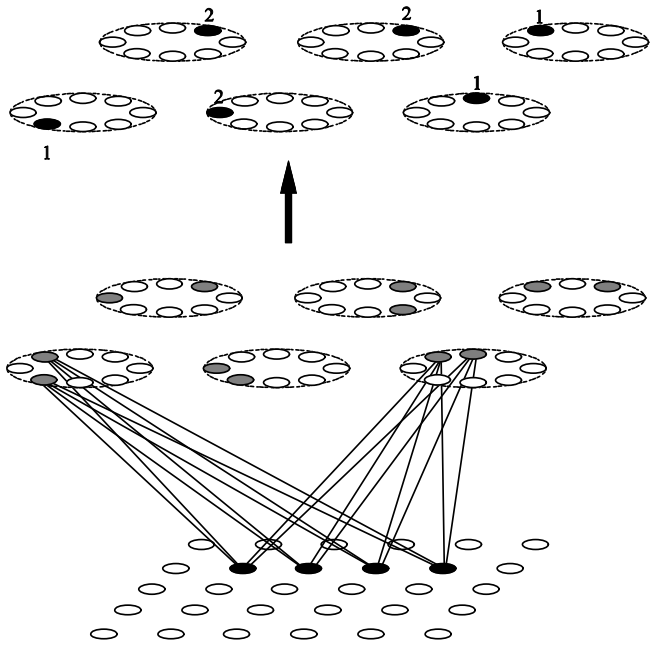

Figure 4.22: Multiple competing hypotheses for an ambiguous prompt. If the model has already experienced and created memory traces for the two sequences, [A B C] and [D B E], the subsequent presentation of the prompt, B, is ambiguous. This figure shows that presentation of the ambiguous prompt, B, leads to the establishment of multiple competing hypotheses (MCHs) in L2. In particular, the model chooses one of the two equally-strongly-implicated L2 cells in each CM with equal probability. Thus, we expect about half of the finally chosen cells to be those from Δ_B^1 ***and about half from*** Δ_B^2***. The finally chosen winners are shown in the top plane. Those from*** Δ_B^1 ***have a '1' above them, and those from*** Δ_B^2***, a '2'. To avoid clutter, only the F- (and thus, R-) connections to two of the CMs are depicted.***

Both dark gray L2 cells in each CM have a ψ value of 4. Since this is, by assumption, an episode-initial time slice, all the dark gray cells have χ values of 1.0, and thus *x* values of 1.0. Thus, the two equally-strongly-implicated cells in each CM are equally likely of being chosen winner. Thus the winner will coincide with Δ_B^1 in about half the CMs and with Δ_B^1 in the other half. ${}^H\theta$ will be halved on the next time slice so that a maximal match will result if either state C or state E presents. For example, suppose state C presents on the next time slice, *t+1*. In this case, the F-vector leads, via Eqs. 4.1 and 4.3, to $\Psi = 1.0$ for the L2 cells in Δ_C^1. At the same time, the

H-vector, carrying only about half as many large (i.e., weight = 1.0) inputs into the cells in Δ^1_C as if the complete Δ^1_B had existed at *t*, still leads, via Eqs. 4.6 and 4.8, to $\Phi = 1.0$ because ${}^H\theta_{t+1}$ has been reduced (in Eq. 4.5) to half its baseline value due to the existence of two MCHs at *t*. Thus, a maximal match condition will exist in each CM at *t+1* and the entire L2 code, Δ^1_C, will be reinstated. Similarly, if state E presents at *t+1,* then Δ^2_E will be fully reinstated. The simulation results reported in Sec. 4.10 show that the model handles MCHs using this method.

The particular situation described in this example forces the addition of a mechanism for controlling the learning rate parameter in the model. Notice that a new internal representation for state B that is a mix of Δ^1_B and Δ^1_B is formed in this example. Denote this novel L2 code as Δ^3_B. This means that if either state C or E present on the next time slice, learning will occur. Specifically, if state C presents, then the H-wts from the L2 cells in $\{\Delta^3_B \setminus \Delta^1_B\}$ onto the cells in Δ^1_C will be increased. This presents a problem given our assumption that there is no new information present in the current instance. That is, the model has previously experienced the sequence [B C] (as a subsequence of Γ^1 = [A B C]) and thus no new information is present in the current instance. Thus, an additional explicit control over the learning rate is needed. Accordingly, the learning rate parameter, η_t, described in Sec. 4.5.13, was added to the model's learning laws. According to Eq. 4.21, $\eta = 0.0$ when either C or E presents in this example, thus no learning occurs.

As discussed in the introduction to this chapter, the general design goal of the model is that as *G* goes to 1.0, the amount of noise added to the winner selection process goes to zero, in which case no new learning occurs. The explicit disabling of learning when $G = 1.0$ can be viewed as a mechanism for handling the special case of MCHs described in this example in a manner consistent with the overall design goals of the model.

Note that alternatively, we could view the presentation of state B in this example as containing new information in the sense that it has never before experienced an input that started with the state B. In this case, one could argue that a new IR for state B should be established and new learning between it and any successor IR is correct. Thus, the mechanism involving explicit control of learning rate adopted herein constitutes a design choice and exploration of alternative design choices is a subject for future research.

4.8.6 Example 6: Ambiguous L2 code

Finally, we consider how the model would handle the two sequences:

Γ^1: [A B]
Γ^2: [A C]

The issues involved in this example are really not so different from those of the previous example. Thus we will provide only a brief explanation. Assume a particular L2 code, Δ^1_A is chosen for the initial occurrence of state A, that another L2 code is chosen for B, and that sequential linkage between the two occurs. Now, when Γ^2_1 = A presents, a perfect match condition will exist and Δ^2_A will be identical to Δ^1_A. That is, Δ^1_A will be fully reactivated. At *t = 2*, there will be essentially complete mismatch between the H-expectancy and the F-expectancy. Thus, noise will completely dominate the winner selection process, resulting in highly unique L2 code for state C.

Thus, the same episode-initial L2 code (for state A) has been equally strongly linked to two different successor L2 codes. Recall of either sequence in the future would require the whole sequences as prompts. In fact, this is psychologically plausible since no human (nor any system in general) could do better in this example.

4.9 Avoiding Saturation

The dependency of noise on similarity, under TEMECOR-II, leads to continuity in the mapping from inputs to internal representations. That is, the more similar two input episodes, X and Y, are, the more similar—i.e., overlapped—their internal representations—i.e., L2 traces—are. The more overlapped two L2 traces are, the larger the number of synapses common to both L2 traces. Thus, assuming X has already been learned, the amount of synaptic increase—i.e., learning—when Y is presented is a decreasing function of overlap between the IRs. This idea was illustrated in Figure 4.1 and is reiterated more specifically in Figure 4.23. Suppose the L2 trace (set of black cells), Δ^X, pictured in panel (a) has been learned previously. The solid lines denote the connections (i.e., the synapses) that would be increased during processing of the input, Γ^X, giving rise to Δ^X. Panel (b) depicts an L2 trace, Δ^Y, which has high overlap with Δ^X. Accordingly, the number of newly

increased synapses, depicted with solid lines, while processing Δ^Y is small. The dotted lines denote connections that were increased while processing Δ^X. Another L2 trace, Δ^W, shown in panel c, has less overlap with Δ^X and thus more learning occurs during processing of Δ^W. Finally, yet another L2 trace, Δ^Z, has even less in common with Δ^X and accordingly, relatively much learning obtains.

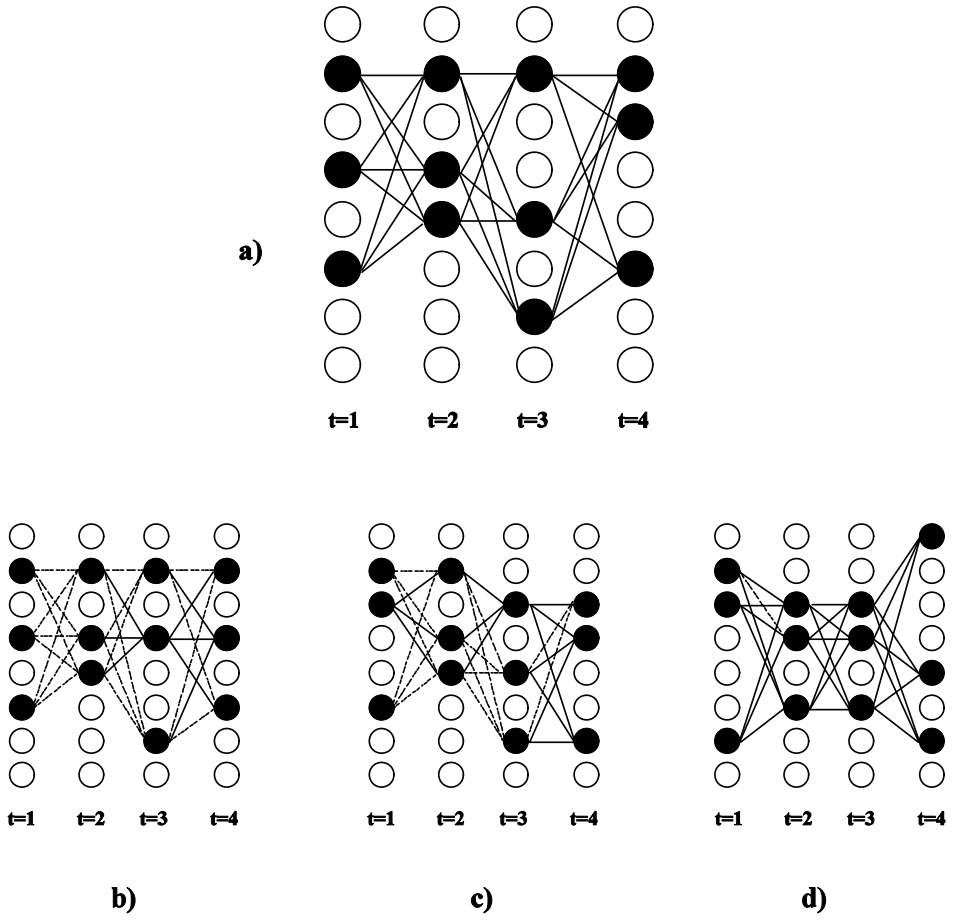

Figure 4.23: The amount of learning increases with decreasing overlap between newly formed and old memory traces. (a) depicts an internal representation—i.e., L2 trace, Δ^X, corresponding to some input, I^X. In order to illustrate the essential point being made here, this figure assumes that L2 contains single cells instead of CMs. The solid lines indicate the connections that would be increased when Δ^X obtains. (b), (c) and (d) show that as we consider L2 traces with progressively less overlap with Δ^X, progressively more learning (i.e., synaptic increase), as indicated by solid lines, occurs. The dotted lines in panels (b), (c) and (d) denote connections that were increased during processing of Δ^X.

Thus the continuity property that TEMECOR-II exhibits provides one mechanism whereby the rate of saturation of synapses is reduced relative to the baseline given by TEMECOR-I. The more familiar the input, the less synaptic increase that will obtain. In fact, this principle suggests that increased correlation over the set of inputs can have a beneficial effect on capacity.

An additional, brute force, method for preventing saturation is to explicitly decrease the degree of plasticity of the synapses as a function of time, or better yet, as a function of the global degree of saturation. This is essentially the method used in Kohonen (1988). Of course, this method has the implication that *no* learning is possible beyond a certain point no matter how novel, interesting, or important subsequent inputs are.

The brute force method becomes interesting when considered in the context of the overall TEMECOR-II model; that is, in terms of the interaction between the horizontal (H) and vertical (F and R) projections. The H-projection contains many more weights than the F-projection. If the number of L1 cells that becomes active to represent an input (i.e., the L1 *coding rate*) is roughly commensurate with or larger than the number of L2 cells that become active to represent the corresponding IR (i.e., the L2 *coding rate*), then the L2 coding rate is a small fraction of the L1 coding rate. This means that the vertical projections should saturate much more quickly than the H-projection. Now suppose the learning in the F-projection is turned off when 50% of the F-wts have been increased. At such time, far less than 50% of the H-wts will have been increased. This suggests that learning in the H-projection be allowed far past the point in time at which learning in the F-projection should be disabled.

This mandatory ‘freezing’ of the vertical projections is used in all the simulations reported herein. If the degree of saturation of the vertical projection is allowed to increase to very high levels—e.g., 70 or 80%, the overall capacity of the model is greatly reduced. The problem concerns episode-initial time slices in particular and the reasoning is as follows. First, note that as L2 codes are always assumed to have S active cells, it is clear that no matter what percentage of F-wts has been increased, no L2 cell can ever have a higher ψ value than S. However, as the number of increased F-wts increases (due to presentation of successive novel inputs), the expected number of L2 cells having $\psi = S$ increases. This trend is graphically depicted in Figure 4.24 (which is based on Figure 3.6). Each panel of this figure depicts a vector of ψ values at the L2 cells of five CMs that would result from presentation of some previously experienced input state. The situation

depicted in panel A corresponds to the case where very few inputs have been presented. Thus, the F-vector clearly picks out one winner in each CM. Panels (b)–(d) show situations corresponding to cases in which progressively larger numbers of inputs are assumed to have been presented. Panel (d), in particular, shows the case where so many inputs have been presented, and such a large percentage of F-wts have been increased, that many cells in each CM are equally strongly implicated. In general, the pool of equally-strongly-implicated cells, on the basis of the F-vector, increases with saturation.

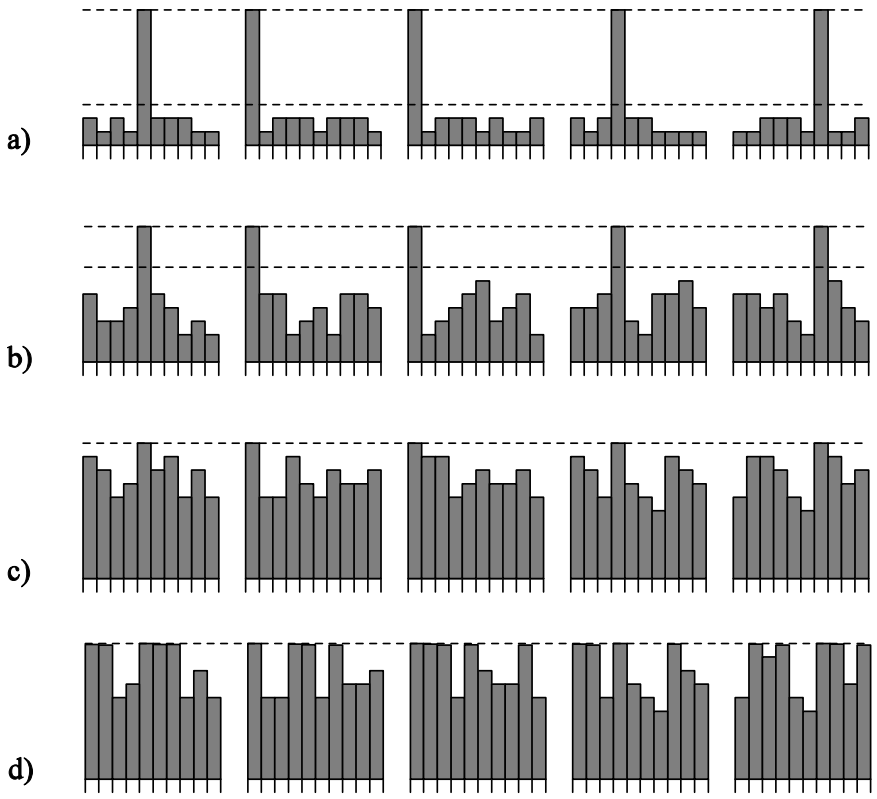

Figure 4.24: (a)–(d) depict the ψ values (i.e., total input via the F-projection) at the L2 cells of five CMs resulting from presentation of some previously experienced input. Progressively larger F-projection saturation values are assumed in going from panel (a) to panel (d). In particular, panel (d) shows the case where so many F-wts have been increased that the F-vector no longer clearly picks out one winner in each CM. The model assumes that the F-projection is frozen (allowing no further weight increases) when about 50% of its weights have been increased. This is currently the limiting factor on the capacity of TEMECOR-II.

Recalling that the rate of saturation of the H-projection is far less than that of the F-projection, it follows that on non-episode-initial time slices, the situation will be as depicted in Figure 4.25. That is, since the H-projection will be far from saturation, the H-vector can be relied upon to clearly pick out one winner in each CM.

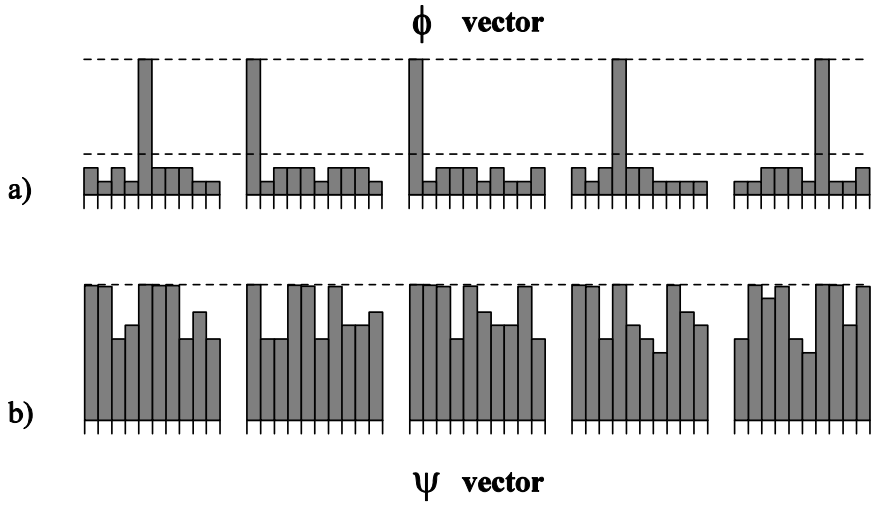

Figure 4.25: The rate of H-projection saturation is much lower than the rate of F-saturation. Thus, even when the F-vector that arises in response to presentation of a previously experienced no longer specifies single winners in each CM, the H-vector will. Of course, the disambiguating influence of the H-vector is only present on non-episode-initial time slices.

However, no H-vector is present on episode-initial slices. Thus, the winner in each CM will be randomly selected from the equally-strongly-implicated pool of potential winners. Thus the number of errors in choosing episode-initial L2 codes increases with F-projection saturation and since incorrect episode-initial L2 codes give rise to erroneous H-vectors on subsequent time slices, a high degree of error will generally obtain on the remaining time slices of the episode. In order to avoid this problem, the model assumes that the F-projection is frozen (allowing no further weight increases) when about 50% of its weights have been increased. This is currently the limiting factor on the capacity of TEMECOR-II.

Determining whether or not the freezing of the forward projection—a constraint that emerges naturally from the model because the horizontal and vertical projections are incommensurate—is

related to the existence of critical periods, which can be viewed as a 'freezing' of weights, would require further research.

4.10 Simulations of TEMECOR-II

The simulation results reported in this section demonstrate that the TEMECOR-II model achieves the various desired properties described in Sec. 4.3. It must be noted that all of the capacity results reported herein are preliminary because very little parameter searching for the purpose of optimization was done.

4.10.1 Preliminary Capacity Result: Solipsistic Recall of Uncorrelated Episodes

Table 4.3 reports the results for a series of simulations illustrating TEMECOR-II's recall capacity. These simulations were performed in the following way. First some number, *E*, of episodes were presented, once each, to the model. All simulations reported in Table 4.3 used uncorrelated episodes consisting of $T = 5$ time slices. Each time slice consisted of $S = 20$ active features chosen randomly from a total pool of $M = 100$ features. Then, to test recall, only the first (i.e., episode-initial) L1 pattern was reinstated, as a prompt. That is, the model formally functioned in the *interactive tracking* mode on the first recall time slice. Then the model shifted to the *solipsistic recall* mode and attempted to output the remaining time slices of the episode based only on the propagation of signals via the H-projection. Thus, the algorithm presented in Sec. 4.7.2 was used on all non-episode-initial time slices during recall. Also, on each non-episode-initial time slice, the L2 code that got reinstated was then used to reinstate the corresponding L1 code. This was done according to Eq. 4.23 and 4.24. That is, the total R-input arriving from the active L2 code was computed for each L1 cell and if the total met or exceeded ${}^{R}\theta$, which was set to 17.5 for all simulations reported in Table 4.3, then that L1 cell was reactivated.

The same recall accuracy measure was used for L2 and L1 and it is identical to that used for L2 in the TEMECOR-I simulations. Specifically,

$$R_{L2}(e) = \frac{C_{L2}(e) - D_{L2}(e)}{C_{L2}(e) + I_{L2}(e)} \qquad (4.25)$$

where $C_{L2}(e)$ is the number of L2 cells that should become active during recall of e, $D_{L2}(e)$ is the number of L2 cells which should have become active but did not (*deletions*), and $I_{L2}(e)$ is the number of L2 cells which should not have become active but did (*intrusions*). The equation for L1 is identical except that the quantities, C, D and *i* refer to L1 cells (features) instead of L2 cells. For completeness the L1 equation is:

$$R_{L1}(e) = \frac{C_{L1}(e) - D_{L1}(e)}{C_{L1}(e) + I_{L1}(e)} \tag{4.26}$$

All simulations also had various model parameters set to maximize the degree of separation between the L2 traces. In particular, the exponent parameter, b, controlling the abruptness of the nonlinearity, Eq. 4.19 was set to 100, thus causing Eq. 4.19 to approximate a step function as in Figure 4.26a below. In addition, the parameters, u, v, and w, that controlled the strength of the nonlinearity of the match functions, Eqs. 4.11 and 4.10, were all set to 10 (see Figures 4.26 b,c). Finally, the exponent parameter in Eq. 4.18 was set to 2. These settings combine to cause the model to tend to choose very different internal representations (i.e., L2 traces) even if the two episodes are very similar. Those model parameters that were constant across all simulations in Table 4.3 are listed in Table 4.2.

Table 4.2: The parameter settings common to all simulations described in table 4.3.

M = 100	*Q = 20*	*K = 50*
S = 20	*T = 5*	
${}^{H}\theta = 11.9$	${}^{F}\theta = 20$	${}^{R}\theta = 17.5$
${}^{\chi}\theta = 0.9$	*b = 100*	*u = 10*
v = 10	*w = 10*	*n = 2*

Table 4.3: Results of simulations of solipsistic recall for uncorrelated patterns. See text for discussion. All Simulations had $^{H}\theta$ ***= 11.9,*** $^{\chi}\theta$ ***= 0.9, S = 20 and T = 5. Abbreviations: E = number of episodes stored;*** $\hat{Z}$ ***= average number of instances of each feature, across entire set of episodes; K = CM size; L = total number of L2 cells;*** $\Omega_{\_H}$ ***= total number of H-weights; H = percentage of H-weights increased;*** $\Omega_{\_F}$ ***= total number of F-weights; F = percentage of F-weights increased;*** R_{set}^{L2} ***= recall accuracy, measured with respect to the L2 traces, over the whole set of episodes; and*** R_{set}^{L1} ***= recall accuracy, measured with respect to the L1 traces, over the whole set of episodes.***

E	*K*	$\hat{Z}$	*L*	*H(%)*	W_F	*F(%)*	R_{set}^{L2}	R_{set}^{L1}
8	10	8	200	38,000	20,000	55.0	94.14	96.52
15	20	15	400	152,000	40,000	53.0	96.83	96.88
22	30	22	600	342,000	60,000	52.0	94.66	98.19
29	40	29	800	608,000	80,000	52.0	93.0	97.65
36	50	36	1000	950,000	100,000	51.0	91.2	96.48
43	60	43	1200	1,368,000	120,000	51.0	89.63	97.2

Table 4.3 shows a linear increase in capacity as *K* increases from 10 to 60. Note that very high accuracy (i.e., > 96%) is achieved at L1 in all cases. Note that rather than present episodes until recall accuracy falls below some criterion (i.e., the testing strategy used for TEMECOR-I), episodes were presented until the degree of saturation, *F*, of the F-projection, reached 50%.

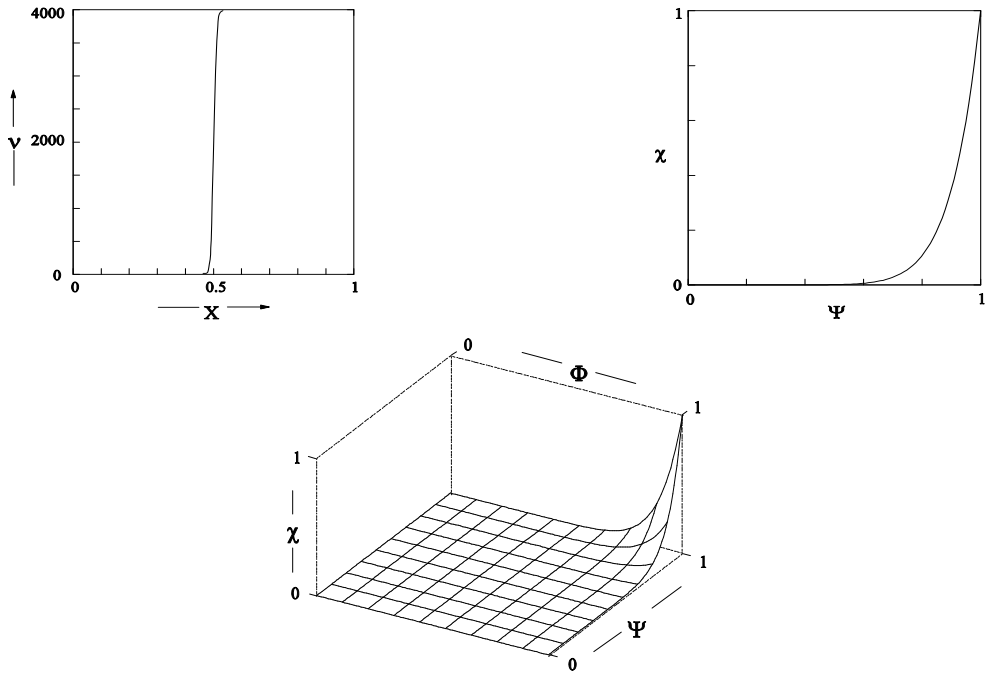

Figure 4.26: (upper left) The function mapping x to ν (Eq. 4.19) with b = 100. (upper right) The episode-initial match function (Eq. 4.10) with w = 10. (lower) The non-episode-initial match function (Eq. 4.11) with u, v = 10.

While the capacity, in absolute terms, is significant, the trend is much slower than that reported in Table 3.2 for TEMECOR-I. More to the point, it is questionable as to whether the demonstrated capacity can be deemed sufficient to explain the apparently very high capacity of human episodic memory. In particular, the capacity is rising linearly while the number of H-weights, W_H, which is the essential measure of network size, increases quadratically. As described in Sec. 4.9, the limiting factor on the capacity of the model is saturation of the F-projection. Most of the storage potential of the horizontal layer remains untapped due to the fact that the F-projection must be frozen when a very small percentage of the H-wts have been increased. For example, in the fourth row of Table 4.3, only 5% of the H-wts have been increased when the F-projection is frozen.

This suggests that some means be found to reduce the rate of saturation of the F-projection. One potential solution is to decrease the amount of learning—i.e., number of weight increases—per time slice in the F-projection as a function of *t*, where *t* is measured from the beginning of each episode. Thus, when *t = 0*, all eligible F-wts (i.e., weights that are currently equal to zero and whose pre- and postsynaptic cells are both active) will be increased. A progressively smaller

fraction of the eligible F-wts would be allowed to increase at *t* increases. This would allow most of the information stored in the F-projection to correspond to the episode-initial and early time slices of episodes—when historical context information (i.e., the H-vector) is absent or less reliable. Recall of later time slices of episodes would be progressively skewed towards increased dependence on the H-vector, whose information becomes more reliable at higher values of *t*. This issue requires further research.

4.10.2 Preliminary Capacity Result: Interactive Tracking of Uncorrelated Episodes

The results reported in this section demonstrate TEMECOR-II's performance when operating in interactive tracking mode. In this mode, the model is simply presented with a series of episodes. The model computes a degree of match, G_t, between its expectation and the actual input on every time slice, *t*. The goal is for the model to reactivate pre-existing L2 traces in proportion to the perceived similarity of the current input in the context of the preceding time slices. Thus, if $G_t = 1.0$, indicating that the current input in the current temporal context is completely familiar, then the model should reactivate the original L2 trace precisely. In contrast, if $G_t = 0$, indicating a completely novel input, the model should attempt to embed an L2 trace as distinct from the set of pre-existing traces as possible. Table 4.4 shows that the model behaves in this fashion.

A total of 16 episodes were presented. The top block of the table corresponds to the initial presentations of episodes. Note that $G = 0.0$ on practically all time slices during these initial exposures to the episodes. Thus, very high noise is added into the winner selection process on all time slices of these first 16 presentations. Accordingly, each episode receives a highly unique trace. These extremely low *G* values reflect the fact that parameters were set, as in the previous simulation (i.e., as in Table 4.2), to bias the system towards mismatch and thus create highly distinct traces, thus maximizing episodic capacity. The bottom block corresponds to the second presentations of each of the 16 episodes. The model correctly registers a perfect match on every time slice of every second presentation, thus virtually no noise is added on these time slices. This shows that the model is treating these second presentations as completely familiar and the nearly perfect recall accuracy—99.14% at L2 and 99.94% at L1—indicate that the original L2 traces are in fact being almost perfectly reinstated.

Table 4.4: Results of one simulation showing the degree of match between the expected and actual inputs, G, computed by the model on each time slice. A total of E = 16 uncorrelated episodes were presented. The model's parameters were set to achieve a maximally dispersed set of memory traces. $^{H}\theta$ ***= 11.9,*** $^{\chi}\theta$ ***= 0.9, S = 20 and T = 5. The match function parameters were set as in the previous simulation (see Figure 4.26). The overall L1 and L2 accuracy were 99.14% and 99.94%. The leftmost column identifies the episode and the other five columns correspond to the five time slices of each episode. The table is broken into two blocks. The second block corresponds to the second presentations of each of the 16 episodes. The model correctly registers a perfect match on every time slice of every second presentation. See text for more discussion.***

Episode	1	2	3	4	5
0	0.00	0.00	0.00	0.00	0.00
1	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00
6	0.02	0.00	0.00	0.00	0.00
7	0.03	0.00	0.00	0.00	0.00
8	0.01	0.00	0.00	0.00	0.00
9	0.06	0.00	0.00	0.00	0.00
10	0.04	0.00	0.00	0.00	0.00
11	0.03	0.00	0.00	0.00	0.00
12	0.13	0.00	0.00	0.00	0.00
13	0.07	0.00	0.00	0.00	0.00
14	0.11	0.00	0.00	0.00	0.00
15	0.12	0.00	0.00	0.00	0.00
0	1.00	1.00	1.00	1.00	1.00
1	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00
6	1.00	1.00	1.00	1.00	1.00
7	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	1.00	1.00	1.00
9	1.00	1.00	1.00	1.00	1.00
10	1.00	1.00	1.00	1.00	1.00
11	1.00	1.00	1.00	1.00	1.00
12	1.00	1.00	1.00	1.00	1.00
13	1.00	1.00	1.00	1.00	1.00
14	1.00	1.00	1.00	1.00	1.00
15	1.00	1.00	1.00	1.00	1.00

A key point this table illustrates is that the model does not need an external signal to tell it whether it is in learning mode or recall mode. If *G = 1.0*, then the original L2 traces are activated,

affording no opportunity for new learning. If $G = 0.0$, highly distinct traces obtain, thus entailing much new learning, as described in Figure 4.23.

Because of the parameter settings used in the previous example, G is very close to zero on practically all the novel episode presentations and equal to one on all familiar episode presentations. However, if parameters are set to achieve the desired continuity property, which will achieve generalization (see Sec. 4.10.5) the model exhibits the property that the amount of learning—i.e., number of increased weights—per time slice is a continuously increasing function of the perceived degree of novelty, G_t. In fact, that property serves as a demonstration of continuity. Figures 4.27—4.29 show, for three different simulations, the amount of learning per time slice, in the F- and H-projections, as a function of the degree of match, G, registered on that time slice. The parameters for the simulation of Figure 4.27 were set so as to achieve maximum continuity and are given in Table 4.5.

Table 4.5: Parameter settings for the simulation maximized for generalization.

$M = 100$	$Q = 20$	$K = 50$
$S = 20$	$T = 5$	
${}^{H}\theta = 13.8$	${}^{F}\theta = 10$	
${}^{\chi}\theta = 0.75$	$b = 2$	$u = 2$
$v = 2$	$w = 2$	$n = 2$

The function shapes are shown in the bottom row of Figure 4.27. All of these parameter settings conspire to bias the model towards sensing relatively higher degrees of match and, thus increasing the degree of continuity in its mappings from inputs to internal representations. This is evidenced by the wider and more gradual spread of the histogram.

The simulation of Figure 4.28 had parameters set so as to achieve a balance in the capacity and generalization capability of the model. This is done by simply increasing the abruptness and curvatures of the x-to-v and match functions as can be seen in the bottom row of Figure 4.28. These parameters are given in Table 4.6.

Table 4.6: Parameter settings for the simulation that achieves a balance between generalization capability and capacity.

$M = 100$	$Q = 20$	$K = 50$
$S = 20$	$T = 5$	
${}^{H}\theta = 13.2$	${}^{F}\theta = 14.3$	
${}^{\chi}\theta = 0.8$	$b = 6$	$u = 4$
$v = 4$	$w = 4$	$n = 2$

Finally, the simulation of Figure 4.29 had parameters set so as maximize capacity. This is done by further increasing the abruptness and curvatures of the x-to-ν and match functions. The parameters for the simulation of Figure 4.29 are the same as those in Table 4.2. The fact that the H-wt histogram shows that all H-wt increases occur when $G = 0.0$ indicates that the model essentially judges any input that has any novelty at all to be completely novel.

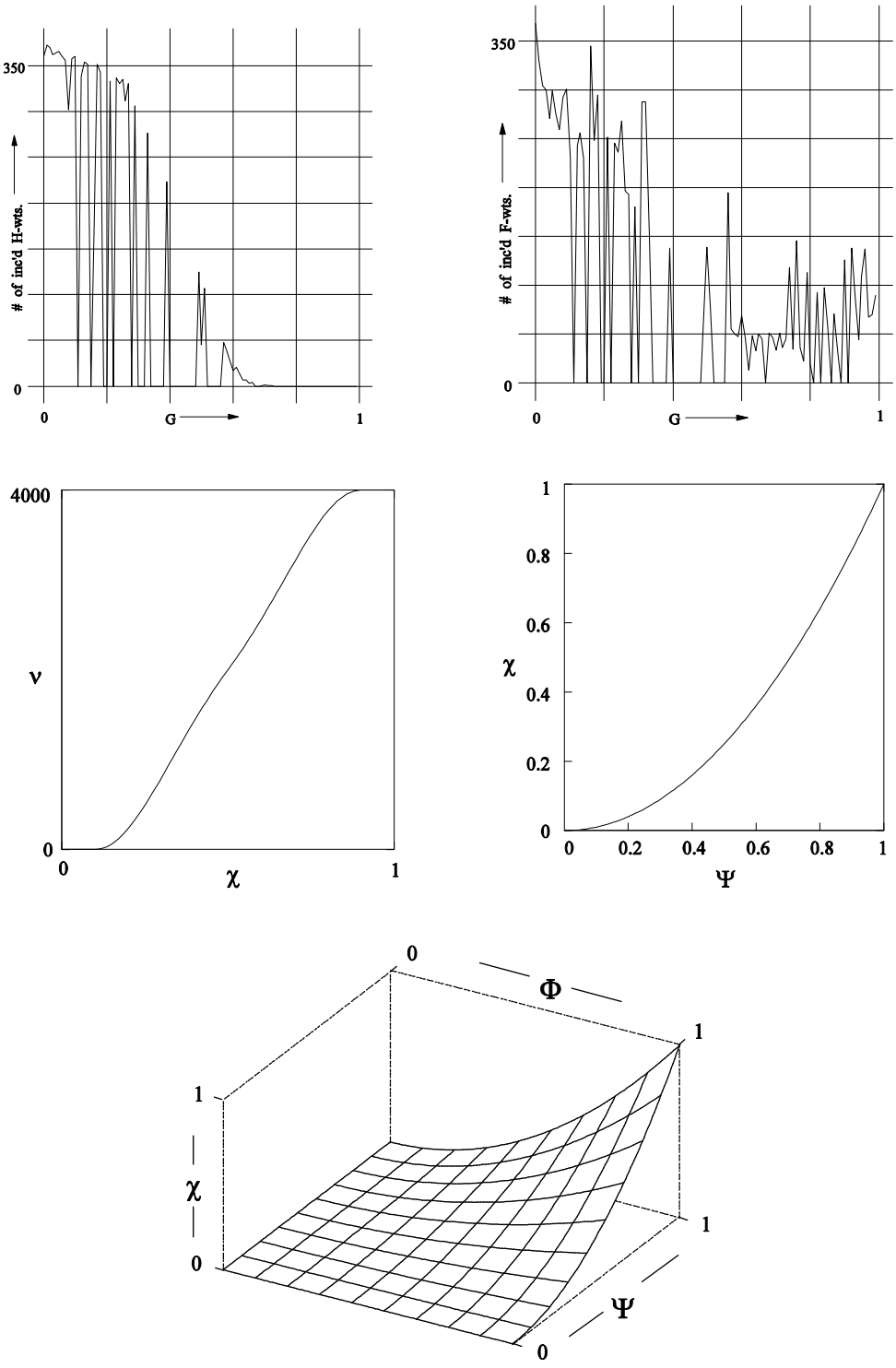

Figure 4.27: (top left) Histogram showing average number of H-wts increased per time slice as a function of G. (top right) Similar histogram for F-wts. (middle left) x-to-ν function where parameter b = 2. (middle right) The match function for episode-initial time slices with w = 2. (bottom) The match function for non-episode-initial time slices with u,v = 2.

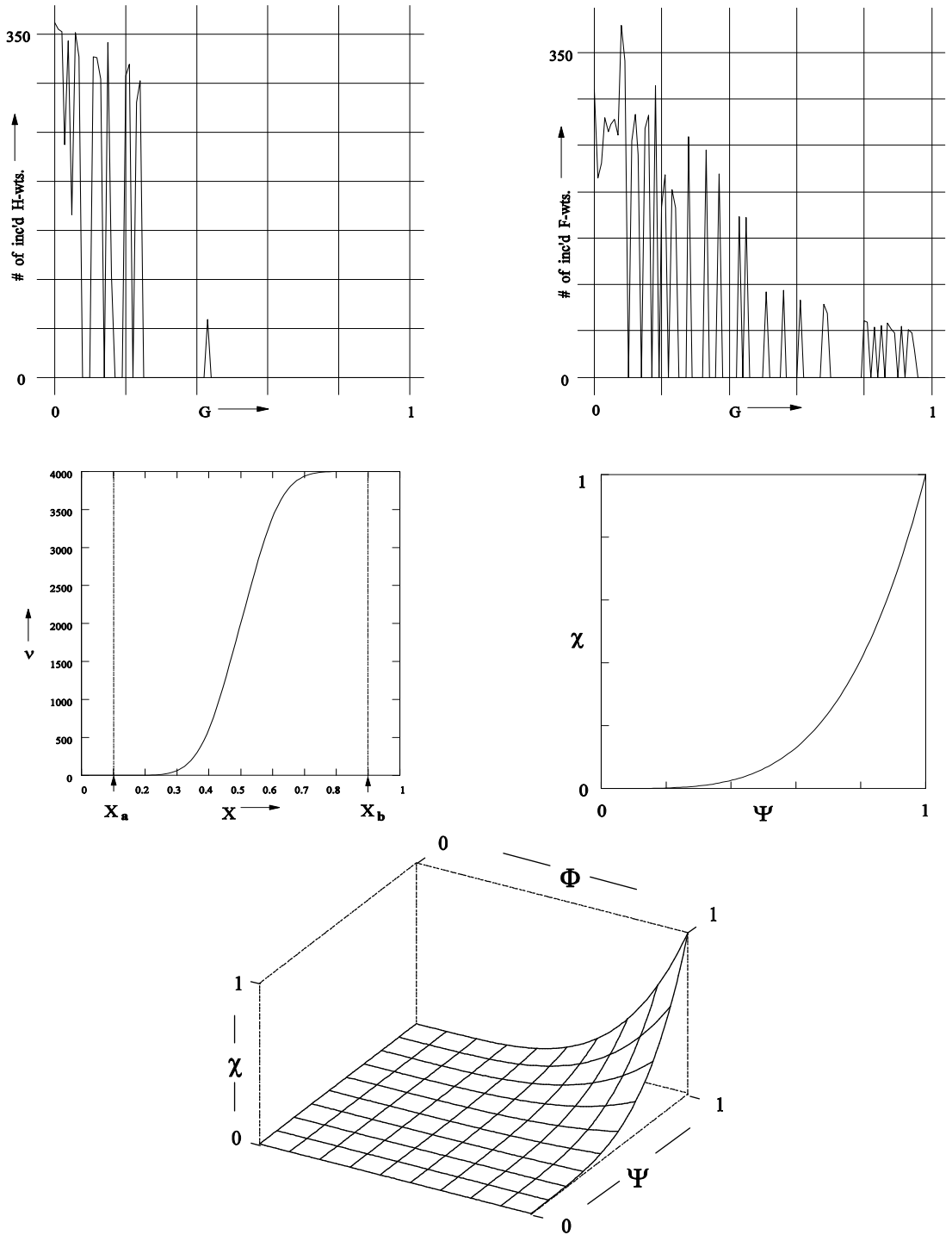

Figure 4.28: (top left) Histogram showing average number of H-wts increased per time slice as a function of G. (top right) Similar histogram for F-wts. (middle left) x-to-ν function where parameter b = 6. (middle right) The match function for episode-initial time slices with w = 4. (bottom) The match function for non-episode-initial time slices with u,v = 4.

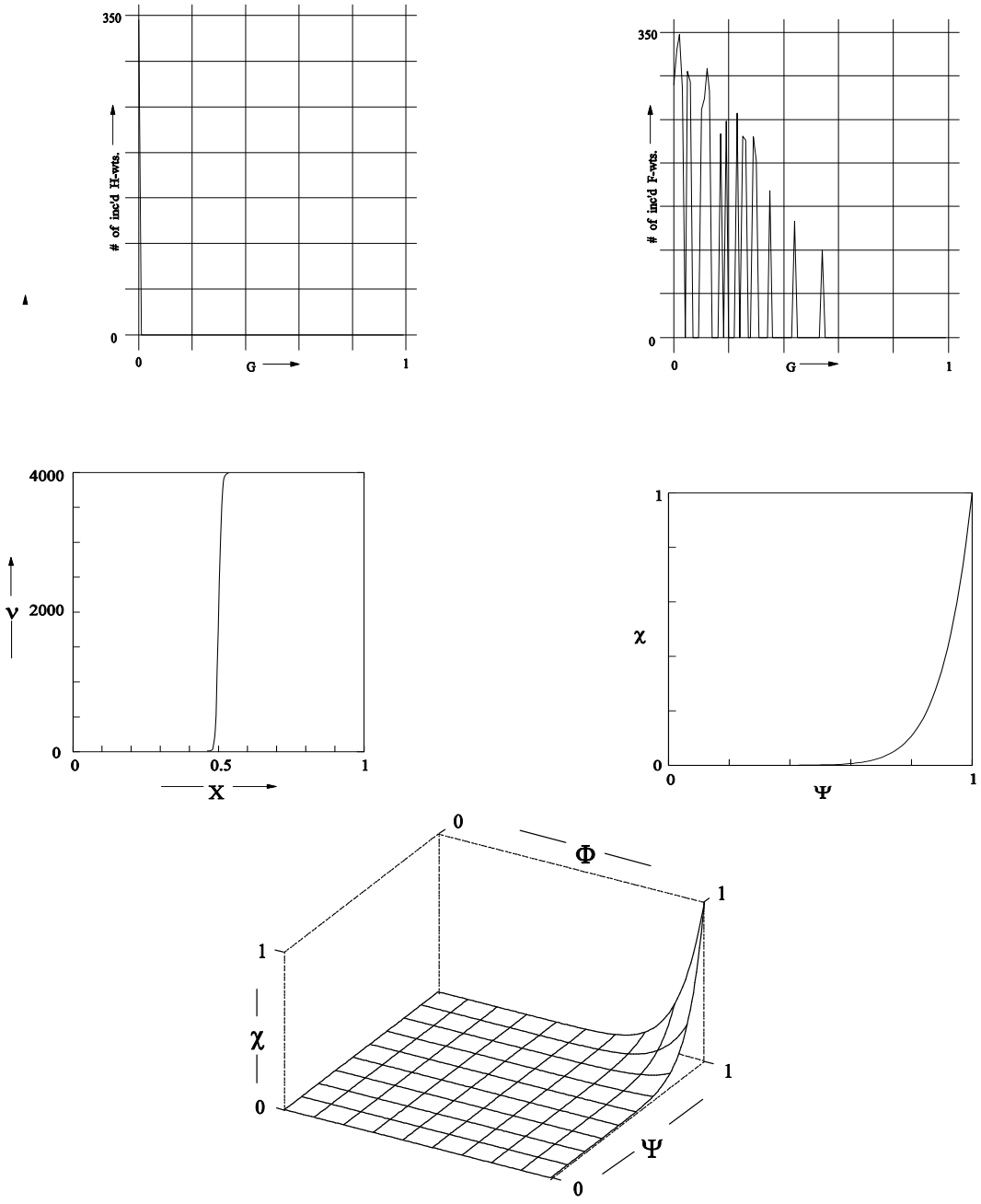

Figure 4.29: (top left) Histogram showing average number of H-wts increased per time slice as a function of G. (top right) Similar histogram for F-wts. (middle left) x-to-ν function where parameter b = 100. (middle right) The match function for episode-initial time slices with w = 10. (bottom) The match function for non-episode-initial time slices with u,v = 10.

4.10.3 Ability to Handle Complex State Sequences and Multiple Competing Hypotheses

The simulation results reported in this section show that TEMECOR-II is capable of processing complex state sequences in the manner described in the algorithm trace section, Sec. 4.8.4.

The first simulation involves presentation of only two sequences:

Γ^1: [B C D]

Γ^2: [D C B]

These episodes are more fully defined as:

Γ_B^1: 1 4 9 12 16 18 25 26 28 32 33 39 48 56 57 58 70 78 87 91

Γ_C^1: 7 16 24 28 36 38 48 50 56 63 66 67 69 74 75 81 85 89 92 94

Γ_D^1: 1 4 6 14 15 29 37 46 51 55 56 60 62 65 67 72 75 84 88 93

Γ_D^2: 1 4 6 14 15 29 37 46 51 55 56 60 62 65 67 72 75 84 88 93

Γ_C^2: 7 16 24 28 36 38 48 50 56 63 66 67 69 74 75 81 85 89 92 94

Γ_B^2: 1 4 9 12 16 18 25 26 28 32 33 39 48 56 57 58 70 78 87 91

Table 4.7 shows actual L2 codes chosen to represent each of the two instances of both episodes. The table also shows the values of G, Ξ and ${}^H\theta$ on each time slice.

The first point to make about these results is that they show that the model can learn a complex sequence set. Table 4.7 shows that Δ_C^1 is very different from Δ_C^2; in fact, they are disjoint. That is, the model finds two completely different internal representations for the state C in the two different presentation contexts. Storing different traces corresponding to the same state is desirable if the goal is to store as many individual episodes as possible. The number of episodes that can be stored and recalled perfectly (i.e., capacity) decreases as the average overlap between L2 traces increases. However, if the goal is to maximize the positive transfer from previous experience to novel instances, then parameters should be set so that the higher overlap between traces is achieved.

Table 4.7: Results of simulation in which the two episodes, Γ^1 and Γ^2, were presented. The top two traces correspond to the initial presentations of the episodes. The last two correspond to the second presentations of the episodes, on which no new learning occurs. G is the degree of match between the expected and actual input. Ξ is the number of multiple competing hypotheses that exist on a given time step. ${}^H\theta$ is the activation threshold that depends on the preceding value of Ξ. See text for discussion.

	L2 Code	G	Ξ	${}^H\theta$
Δ^1_B	12 20 **9** **24** 25 **16** 10 18 **21** 19 **4** 39 12 13 6 17 **33** 9 **37** 33	0	0	16
Δ^1_C	31 4 22 14 12 13 17 0 4 14 14 17 34 24 2 19 0 12 38 22	0	0	16
Δ^1_D	5 25 20 34 31 4 14 2 38 3 29 20 10 15 33 31 2 3 6 1	0	0	16
Δ^2_D	5 25 20 34 31 4 14 2 38 3 29 20 10 15 33 31 2 3 6 1	1	1	16
Δ^2_C	32 25 11 23 29 0 11 25 7 35 17 32 27 12 11 23 1 38 35 36	0	0	16
Δ^2_B	27 0 4 5 8 33 17 37 8 0 37 11 14 34 1 18 0 0 15 31	0	0	16
Δ^1_B	27 0 **9** **24** 8 **16** 17 37 **21** 0 **4** 39 12 13 1 18 **33** 0 **37** 31	1	2	16
Δ^1_C	31 4 22 14 12 13 17 0 4 14 14 17 34 24 2 19 0 12 38 22	1	1	8
Δ^1_D	5 25 20 34 31 4 14 2 38 3 29 20 10 15 33 31 2 3 6 1	1	1	16
Δ^2_D	5 25 20 34 31 4 14 2 38 3 29 20 10 15 33 31 2 3 6 1	1	1	16
Δ^2_C	32 25 11 23 29 0 11 25 7 35 17 32 27 12 11 23 1 38 35 36	1	1	16
Δ^2_B	27 0 4 5 8 33 17 37 8 0 37 11 14 34 1 18 0 0 15 31	1	1	16

The second point to make is that when the sequences are presented for a second time, the model recognizes these second presentations as familiar. Precisely the correct L2 codes are reinstated during the second presentations of the episodes, except for the first time slice of the second presentation of Γ^1. This is an instance of multiple competing hypotheses (MCHs). In fact, the model behaves correctly in this case. The reasoning is as follows. The model has previously chosen two completely different L2 codes, Δ^1_B and Δ^2_B, for the two instances of state B, Γ^1_B and Γ^2_B. The same set of L1 cells, Γ_B, is linked equally strongly to these two different L2 codes. When state B is presented again at the beginning of the second presentation of Γ^1, the two different cells in each CM, one from Δ^1_B and the other from Δ^2_B, are equally strongly implicated—i.e., have equal ψ values. Since it is an episode-initial time slice, the degree of match depends only on the F-inputs. Thus, $\chi = 1$ for all of these cells. This leads to equal probability of picking either cell in each CM.

We expect the final winner to be the cell from Δ_B^1 in about half of the $Q = 20$ CMs and the cell from Δ_B^2 in the other half. In fact, this is exactly the outcome as can be seen in the table. The 10 cells contained in the resulting L2 code for the second instance of Γ_B^1 that are in common with those from the original instance of Γ_B^1 are bolded. The reader can check that the other cells are common to the initial instance of Δ_B^2. As explained in Sec. 4.8.5, $G = 1.0$ implies that $\eta = 0.0$ in this case and so no new learning occurs.

As described in Sec. 4.5.3, the existence of two MCHs on this time slice implies that the set of L2 cells, Δ_C^1, that should become active on the next time slice, when state C presents, will have ϕ values of 10 which is far below ${}^H\theta_{baseline} = 16$. However, since two MCHs currently exist ($\Xi = 2$), ${}^H\theta$ is halved on the subsequent time slice, as can be seen in the table. This allows a perfect match ($G = 1.0$) to be registered on the next time slice, and the *entire* L2 code, Δ_C^1, to be reinstated. A single hypothesis remaining, the rest of the L2 trace then reads out correctly—i.e., tracks the input correctly.

4.10.4 Demonstration of correct recall for a larger CSS set

This section describes the results of a simulation involving the set of $E = 7$ CSSs listed in Table 4.8. The overall accuracy achieved for this simulation was 83.03% at L2 and 91.43% at L1. Other parameters for this simulation are given in Table 4.9.

Table 4.8: The set of episodes used in this simulation.

Γ^1:	[T W Q S V]
Γ^2:	[F L W T D]
Γ^3:	[P D M E R]
Γ^4:	[X Y T C C]
Γ^5:	[S G W L I]
Γ^6:	[H D T U X]
Γ^7:	[P W N X Y]

Table 4.9: The parameter settings for simulation involving episodes of table 4.8.

$M = 100$	$Q = 20$	$K = 50$
$S = 20$	$T = 5$	$U = 25$
${}^{H}\theta = 12.5$	${}^{F}\theta = 20.0$	${}^{R}\theta = 17.0$
${}^{\chi}\theta = 0.9$	$b = 5$	$u = 4$
$v = 4$	$w = 10$	$n = 2$

The main point of this simulation is to demonstrate that the model is capable of storing and recalling a set of complex sequences with high overall accuracy. It can be seen by comparing these results to those reported in Table 3.3 that TEMECOR-I exhibits far greater capacity than TEMECOR-II. The same explanations regarding the smaller capacity of TEMECOR-II that were made at the end of Secs. 4.9 and 4.10.1 apply here. Again, these are preliminary results intended to show that the model is viable.

Examination of the slice-by-slice values of G computed by the model, shown in Table 4.10, reveals that the model recalls five of the seven episodes virtually perfectly. It fails when attempting to recall Γ^1 and Γ^4. The problem with Γ^1 is that it begins with state T that occurs four times over the whole input set. When T is presented as a prompt—i.e., the second presentation of Γ^1, the model cannot be sure which of the four instances of T is presenting. Thus it must allow $\Xi = 4$ MCHs to become active in L2. Thus, the correct winner—i.e., the one that was active in the initial presentation of Γ^1_t—is expected to again become active in approximately Q/4 = 5. Because $\Xi = 4$, ${}^{H}\theta$ is set to 1/4th its baseline value—i.e., 3.12—on the next time slice, when state W presents. The problem is that due to the statistical nature of the model, only 3 L2 cells from the initial instance of Γ^1_t are reinstated. This is less than ${}^{H}\theta$, and this eventually leads to a very low G value, which subsequently leads to a highly unique L2 code being chosen for this second time slice (and the remaining three as well) of this second instance of Γ^1.

Table 4.10: The values of G computed by the model on each time slice. The rightmost column in the lower block gives the length of the prompt, φ—that is, the number of time slices used to prompt the model to recall the sequence.

Episode	*1*	*2*	*3*	*4*	*5*	*φ*
1	0.00	0.00	0.00	0.00	0.00	
2	0.00	0.00	0.00	0.00	0.00	
3	0.00	0.00	0.00	0.00	0.00	
4	0.00	0.00	0.00	0.00	0.00	
5	1.00	0.02	0.00	0.00	0.00	
6	0.00	0.00	0.00	0.00	0.00	
7	1.00	0.07	0.00	0.00	0.00	
1	1.00	0.15	0.00	0.00	0.00	2
2	1.00	1.00	1.00	1.00	1.00	2
3	1.00	1.00	1.00	1.00	1.00	3
4	1.00	1.00	0.00	0.00	0.00	2
5	1.00	0.90	1.00	1.00	1.00	2
6	1.00	1.00	1.00	1.00	1.00	2
7	1.00	1.00	1.00	1.00	1.00	3

Recall of Γ^4 fails because, by the end of the prompt period, in which it has been presented with the first two states of Γ^4, [X Y], the model has mistakenly *locked into* the L2 trace sequence corresponding to the last two time slices of Γ^7 (which is another instance of [X Y]) in the data set.

The model is statistical in nature and will make errors like these with probability depending on the model parameters and statistics of the input set. One possible means of reducing this problem is to make the internal representation (L2 code) chosen at *t* depend on some small window of previous L2 codes. That is, we could make the model's dynamics less local in time by generalizing the H-projection so that it contains connections of various temporal delays. This general technique of explicitly *concentrating* context in time forms the basis of many other proposals as discussed in Ch. 2.

4.10.5 Generalization Results

The final simulation results provide direct evidence of the generalization property, as described in Sec. 4.3.2. All four simulations described in this section involved uncorrelated episodes. Each column of Table 4.11 gives the parameters for one of the simulations.

Table 4.11: Each column gives the parameter set corresponding to one of the simulations.

Parameter	*1*	*2*	*3*	*4*
M	100	100	100	100
Q	20	20	20	20
K	50	50	50	50
S	20	20	20	20
T	5	5	5	5
${}^{H}\theta$	16.0	13.6	16.0	16.0
${}^{F}\theta$	20.0	10.0	10.0	10.0
${}^{\chi}\theta$	0.9	0.8	0.75	0.75
b	100	11	5	5
u	10	5	3	4
v	10	5	3	3
w	10	7	4	4
n	2	2	2	2

The simulations were performed as follows. E episodes were presented, once each. Then, perturbed versions, differing by $d = 2, 4, 6$, or 8 (out of 20) features per time slice from the original episodes were generated. Then the model was tested by presenting the first φ time slices of the perturbed episodes as prompts. Following the prompt time slices, the model entered solipsistic mode (i.e., cutting off any further input) and processing continued, from that point, merely on the basis of signals propagating in the H-projection. The results indicate that the model was extremely good at locking into the trace corresponding to the most-closely-matching original episode. One of the episodes, Γ^1, from this simulation, and its perturbed version, Γ^{1*}, are listed below. Prompts consisted of two time slices in this simulation (see Table 4.12). Those features which differ between the corresponding slices of Γ^1 and Γ^{1*} are bolded. Note that the first two time slices of the original episode are quite different from the first two time slices of the perturbed episodes. In fact, by Eq. 4.26, which is used to measure the similarity between two episodes, the original time slices (in this simulation) are only 66.6% similar to the corresponding perturbed time slices.

Γ_1^1 : 1 **6** 12 18 27 31 **34** 38 40 42 47 49 57 77 79 80 81 **82** 96 **98**

Γ_2^1 : 3 6 21 22 **25** 26 31 **33** 36 42 43 **53** 58 63 69 73 74 79 **87** 95

Γ_3^1 : 0 5 10 19 20 29 31 35 40 41 51 52 58 64 73 77 82 84 85 94

Γ_4^1 : 3 6 9 14 15 22 28 32 36 39 44 46 59 63 66 67 79 84 85 86

Γ_5^1 : 0 1 2 15 21 22 24 28 39 41 43 44 49 54 67 68 72 73 77 88

Γ_1^{1*} : 1 12 18 27 31 38 40 42 47 49 57 **59** 67 74 77 79 80 81 **83** 96

Γ_{21}^{1*} : 3 6 21 22 **24** 26 31 36 **41** 42 43 **50** 58 63 69 **70** 73 74 79 95

Γ_3^{1*} : 0 5 10 19 20 29 31 35 40 41 51 52 58 64 73 77 82 84 85 94

Γ_4^{1*} : 3 6 9 14 15 22 28 32 36 39 44 46 59 63 66 67 79 84 85 86

Γ_5^{1*} : 0 1 2 15 21 22 24 28 39 41 43 44 49 54 67 68 72 73 77 88

As Table 4.12 shows, the model *locks into* the L2 trace corresponding to the most-closely-matching original episode (i.e., the episode from which the novel one was created by perturbation). The accuracy measure in the table measures how close the current L2 trace is to the L2 trace of the most-closely-matching original episode. Very high accuracy is achieved in all four simulations. The accuracy reported in the last line of the table (82.73%) may seem low. However, if the accuracy measure is taken only for the final time slice then it is close to 100% for all four simulations. The view taken herein is that given that the stimuli being presented are spatiotemporal in nature, the most relevant measure of performance is the measure of accuracy on the last time slice of the test episode. That is, if the model can lock into the correct memory trace by the end of the test trace, then that should be taken as sufficient evidence that model has recognized the current perturbed episode as an instance of a familiar episode.

Table 4.12: Row i gives the results for simulation i. These four simulation results demonstrate the model's ability to generalize—i.e., to treat similar inputs in a similar way.

Simulation	**E**	***d***	φ	R_{set}
1	27	2	1	92.3%
2	13	4	1	98.0%
3	7	6	1	98.3%
4	13	8	2	82.7%

It is particularly instructive to look at the simulation from the fourth row of Table 4.12 in more detail. Table 4.13 gives the L2 recall accuracy for each time slice during presentation of each of the perturbed episodes. It can be clearly seen that on all but one of the test trials, the model locks into the correct L2 trace over the course of the two-slice-long prompt so that by the third time slice, when the model enters solipsistic mode, the L2 accuracy measure is 100%.

Table 4.13: Per-Time-Slice L2 Accuracy for the Test Trials of Simulation 4 of Table 2

Episode	***T = 1***	***T = 2***	***T = 3***	***T = 4***	***T = 5***
1	0.9	0.9	1.0	1.0	1.0
2	0.82	1.0	1.0	1.0	1.0
3	0.67	1.0	1.0	1.0	1.0
4	0.82	0.9	1.0	1.0	1.0
5	0.67	0.82	1.0	1.0	1.0
6	0.67	0.9	1.0	1.0	1.0
7	0.9	1.0	1.0	1.0	1.0
8	0.74	1.0	1.0	1.0	1.0
9	0.74	1.0	1.0	1.0	1.0
10	0.67	0.82	1.0	1.0	1.0
11	0.54	0.67	0.22	0.0	0.0
12	0.48	0.21	0.0	0.0	0.0
13	0.82	0.9	1.0	1.0	1.0

These simulations provide preliminary evidence that TEMECOR-II exhibits generalization, and in fact categorization, in the spatiotemporal domain. Furthermore, these simulations spanned a wide range of parameter settings. Even for the simulation for which $d = 2$, in which parameters were set to maximize capacity and therefore should lead to less generalization capability, the model is able to lock into the appropriate traces virtually perfectly.

4.11 Relation to Work of Hasselmo and Colleagues

As explained in Sec. 4.1, in particular with reference to Figure 4.1, the effect of prior learning, manifest in the pattern of modified weights in an associative matrix, is generally to reduce pattern

separation. Analyses supporting this same conclusion for a similar, although spatial, associative model are presented in O'Reilly & McClelland (1994). Hasselmo (1993, 1994) has previously pointed out that many popular abstract associative memory models to date (Amit, 1988; Anderson, 1972; Hopfield, 1984; Kohonen, 1972, 1988; Palm, 1980), avoid this problem by simply assuming complete suppression of transmission via the modifiable synapses during learning. On the other hand, during recall trials, transmission via the matrix is fully effective so that the correct, previously learned, associations can be recalled. Although this assumption of two different dynamics—one for learning, one for recall—has previously gone unjustified from a neurobiological standpoint, Hasselmo et al. (1991, 1992) provide neurophysiological evidence for a specific mechanism involving acetylcholine (ACh) for controlling transmission via the modifiable matrix.

They have studied rat piriform cortex in which the recurrent *intrinsic* projection, analogous to the modifiable associative matrix, is confined to layer Ib and is physically segregated from the *afferent* projection, which is confined to layer Ia. Both projections impinge on the same set of cells. Specifically, Hasselmo and his colleagues have found that increased presence of acetylcholine (ACh) suppresses transmission in the intrinsic but not afferent synaptic projections in piriform cortex.

Their proposal is that when novel patterns are present, the system's dynamics should be set for learning and in order to assure that prior learning, present in the intrinsic matrix, interferes minimally with existing traces, the ACh level should be high. In contrast, when a familiar pattern is present, dynamics should be set for recall. In this case, the influence of the intrinsic matrix is needed in order to cause the trace laid down during the initial experience of the input to read out; thus, the ACh level should be low. This proposal suggests that the level of cortical acetylcholine, which is supplied by the basal forebrain, be controlled by the overall level of novelty detected by the system. Electrode studies of the basal forebrain of behaving animals (Wilson & Rolls, 1990) support the possibility that ACh level is modulated as a function of novelty of input.

TEMECOR-II utilizes the same general functionality achieved by the ACh mechanism proposed by Hasselmo; i.e., controlling the relative influence of two classes of input to a given population of cells. The major difference is that in Hasselmo's model, both inputs—i.e., the intrinsic matrix and the afferent matrix—are deterministic, whereas in TEMECOR-II, one input is deterministic—i.e., the combined F- and H-input vectors—and the other is noise.

4.12 Weaknesses of TEMECOR-II

Although TEMECOR-II exhibits many of the functional properties of episodic memory and generalization, it has the following two problems:

a) In order to exhibit psychologically realistic memory capacity, it requires a neurobiologically unrealistic degree of intrinsic connectivity, specifically within entorhinal cortex (EC). (See Figure 4.30)

b) It does not explain the temporal gradient of retrograde amnesia (Ribot, 1882; Squire, Cohen & Nadel, 1984).

A speculative hypothesis addressing both of these problems follows in the next section.

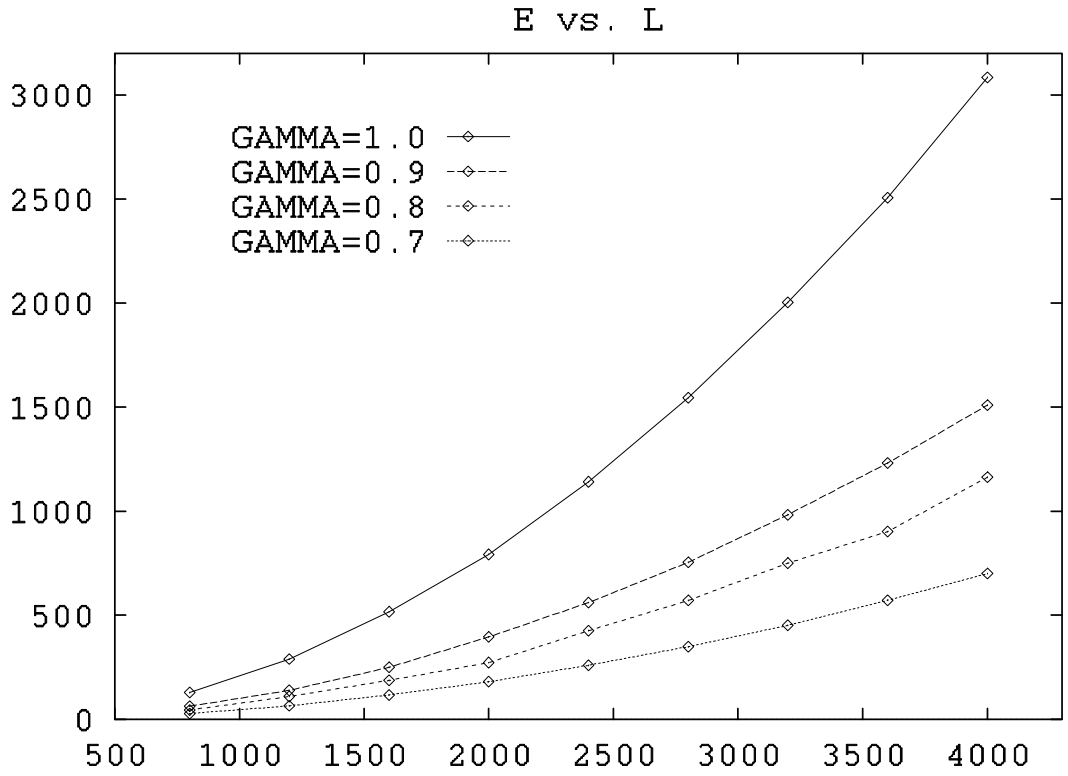

Figure 4.30: Although the qualitative faster-than-linear relationship between E (y-axis) and L (x-axis) is preserved as the degree of horizontal connectivity, γ (GAMMA), is reduced, the rate of increase drops off rather quickly. Given that actual horizontal (intrinsic) connectivity rates in cortex, specifically entorhinal cortex, are probably at most 2-3% (within patches encompassing on the order of hundreds of mini-columns), the theory requires some neurophysiologically plausible means of effectively boosting the interconnectivity rate. This is the major hypothesized function of the hippocampal complex within the TEMECOR modeling framework.

4.13 Hypothesis Regarding Function of Hippocampus

In TEMECOR-II, the H-vector arising from Δ_t interacts with the F-vector at t+1 to determine Δ_{t+1}. Then Hebbian learning takes place between Δ_t and Δ_{t+1}.[23] This mechanism assumes a) physical connectivity between the L2 cells in Δ_t and Δ_{t+1} and b) that the weight increases due to a single instance of $[\Delta_t, \Delta_{t+1}]$ are sufficient to reliably reinstate Δ_{t+1} following the next instance of Δ_t. As noted in the previous section, the first assumption holds only over very small regions of cortex. If the theory is to account for the establishment of reliable linkages across larger spans of cortex, then it must be generalized to function in the context of multi-synaptic pathways. If the second assumption is false, then some other mechanism, besides the vector of signals propagating in the H-projection, is needed in order to cause reliable reactivation of sequences of L2 codes based only on the initial instance of such sequences. This section outlines a speculative proposal in which the hippocampal loop that leads from entorhinal cortex (EC) back, via the stages of the hippocampus, to EC, addresses both problems. As it turns out, certain ancillary assumptions of this hippocampal hypothesis also suggest a possible explanation of the temporal gradient of retrograde amnesia that is generally similar to various other proposals in which the hippocampus is viewed as training the cortex (McClelland et al., 1994; Murre, 1995)—except that these other models are defined only for the domain of spatial patterns.

We have stated earlier that the model's L2 is considered to be analogous to the entorhinal cortex (EC). Thus, in broad terms, the hypothesis is that each EC representation (L2 code) simultaneously gives rise to two paths of neural excitation: one via the intra-EC (i.e., intrinsic) connectivity matrix—that is, the H-projection, and one via the hippocampal loop, as suggested in Figure 4.31. In keeping with the nomenclature used so far, let the sequence of EC codes depicted in the Figure 4.31a be called Δ^1. Notice that Δ^1 is represented as beginning at $t = 2$. This is to facilitate comparison with the temporally overlapped sequence, Δ^2, shown in panel b of the figure. Learning is assumed to occur in both the H-projection and the hippocampal projection (HIPP-projection). However, in keeping with neurobiological evidence documenting the rapid and robust establishment of hippocampal traces (Wilson & McNaughton, 1994), we assume the rate of learning is much higher in the HIPP-projection than in the H-projection. Thus, we assume that after

---

[23] Λεαρνινγ αλσο τακεσ πλαχε φρομ Γ_{t+1} to Γ_t, however that is not the focus of this section.

one instance of Δ^1, Δ^1_2 can, via the HIPP-projection, reliably cause Δ^1_5 to become active three time steps later. In contrast, the learning rate of the H-projection is assumed to be too small to allow, after only a single previous instance, reliable reinstatement purely on the basis of the H-projection. However, with each successive hippocampally-mediated reinstatement of a Δ sequence, the H-wts can be increased, so that eventually, the H-projection would support reinstatement of the sequence without the help of the hippocampus. As with the other theories cited above, this general scenario explains the existence of the retrograde amnesia gradient. If the hippocampus was suddenly removed, then the oldest memories, which no longer rely on the hippocampal traces, would be preserved, the most recent memories, relying most heavily on the hippocampus, would be lost, and there would be a retrograde gradient of loss in between.

Figure 4.31 also suggests a solution to the connectivity problem mentioned above. Specifically, even if we assume that the direct connectivity of individual EC cells is rather limited (e.g., to a radius of 5 mini-columns), the time delay introduced by the multiple stages of the hippocampus would, in principle, allow the linking mechanism described here to establish multi-synaptic pathways across substantial regions of EC. Figure 4.32, taken from Levy (1989), explicitly shows that any cell in EC can be reached, via the multiple divergent stages of the hippocampus, from any other EC cell. Thus, the fundamental connectivity required by the hypothesis exists. Whether the intra-EC connectivity and the various timing requirements of a more detailed instantiation of this hypothesis are consistent with neurobiology is a question for future research.

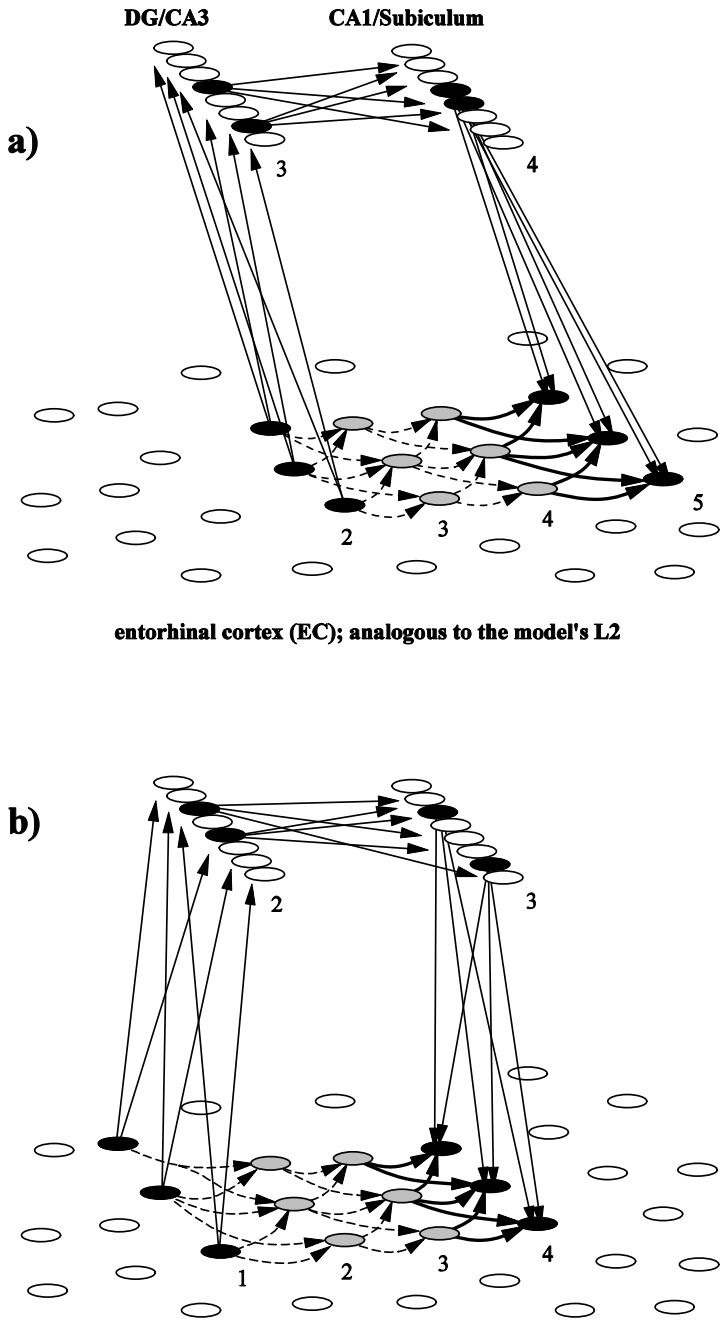

Figure 4.31: Sketch of the hypothesis, described in this section, in which the hippocampus solves various potential connectivity and learning rate problems of the current version of the theory. Note that the hippocampal loop is idealized as having two stages: DG/CA3 and CA1/subiculum. Note that this figure does not depict the fact that EC is hypothesized to be composed of competitive modules (CMs). All weights (depicted as arrowheads in this figure) are plastic. See text for explanation of figure.

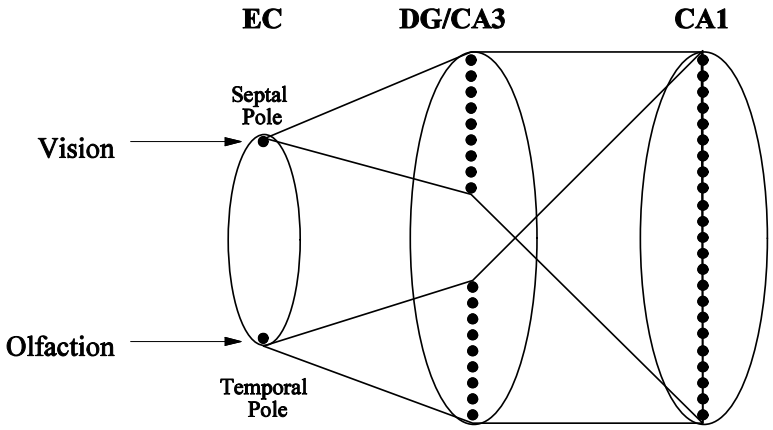

Figure 4.32: This figure shows that a path exists from any point along the entire septo-temporal extent of the EC to any point along the entire length of CA1. Given the projections from CA1/subiculum back to EC, it follows that signals from any cell in EC can reach any other EC cell via a multi-synaptic path through the hippocampal loop. This figure is redrawn from the bottom diagram of Fig. 3 of Levy (1989). Copyright ©1989 by William B. Levy.

Note that in this scenario, the hippocampus is viewed as supplying a training input to EC. However, as suggested by Figure 4.31, the linkage being directly trained is not between Δ_2^1 and Δ_5^1, but rather between Δ_4^1 and Δ_5^1. This is highlighted by the use of bold solid arrows between Δ_4^1 and Δ_5^1 in the figure. As Figure 4.31b suggests, Δ_4^1 may be assumed to be reliably reinstated by a similar vector of activation in the HIPP-projection arising from Δ_1^1. Thus, the hypothesis is that immediately following the initial occurrence of any particular Δ sequence, that sequence can be completely reinstated by these temporally overlapped activations of hippocampal mappings.

Note however that immediately following the initial instance of Δ^2, reinstatement of Δ_1^2 can be relied upon to cause Δ_4^2, but not the intervening EC representations, Δ_2^2 and Δ_3^2, because that relies on the H-wts which remain very weak after a single trial. Similarly, Δ_4^2 can be relied upon to cause Δ_7^2, and Δ_7^2 can be relied upon to cause Δ_{10}^2, etc. This state of affairs is suggested in panels a–c of Figure 4.33. However, in order to reliably reinstate all time slices of a Δ sequence immediately following its first occurrence, a spatiotemporal prompt encompassing the first k time slices of the sequence, where k is the delay through the hippocampal loop, is required. This is schematized in panel Figure 4.33d. Eventually, with enough trials, the intra-cortical traces become sufficiently

strong that perhaps only the first time slice would be needed to reinstate the entire sequence. The figure also depicts the fact that in general, different representations within the various fields of the hippocampus will arise in response to different input EC patterns.

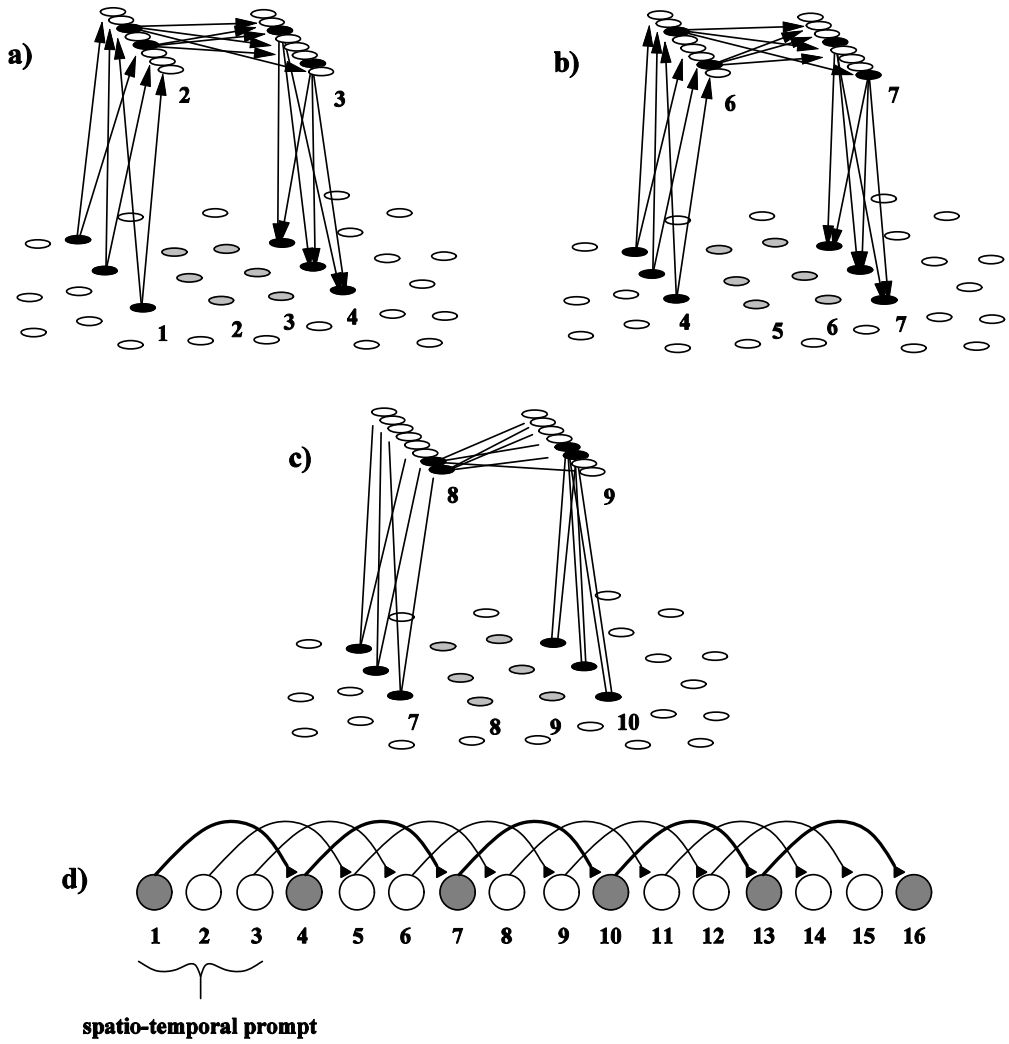

Figure 4.33: Following one instance of a Δ sequence, reinstatement of a particular L2 code—e.g. Δ_1 in panel a—reliably reactivates only the L2 code that occurs k time steps later, where k is the hippocampal delay—e.g. Δ_4. It does not reliably reactivate the intervening L2 codes, Δ_2 and Δ_3 because that involves the H-wts that are still very weak after only one trial. Panels b and c show the two next reliable reactivations in this thread of the trace. Until the H-wts are sufficiently strong, reliable reactivation of all time slices comprising the whole trace requires a spatiotemporal prompt that includes the first k time slices of the sequence, as schematized in panel d.

Murre (1995) describes a theory which also calls upon the hippocampus to solve the *connectivity problem* which in Murre's terms is essentially as follows: while the intrinsic connections of cortex are, for the most part, highly localized, our memories are composed of multi-modal elements presumably involving highly distant neural codes. How can these distant neural codes be reliably and quickly linked? This is related to the 'binding problem' (Engel, Konig, Kreiter, Schillen & Singer, 1992). Given our interpretation of TEMECOR's L2 as analogous to EC, and the fact that EC receives inputs from numerous brain regions representing all modalities, the connectivity problem reduces to finding a way to boost the effective interconnectivity of EC.

Figures 4.31 and 4.32 suggest that the projection via the hippocampal loop boosts the *effective* degree of EC connectivity. However, due to its limited size, it cannot retain mappings indefinitely. If it did not lose information, through either passive or active decay, over some time course, the projection would eventually become saturated (i.e., all weights would grow large) and it would be useless. Regarding this, there is evidence for long-term depression (LTD) in both the mossy fibers and the Schaffer collaterals (Levy & Desmond, 1985; Levy et al., 1990). Therefore, the hippocampus could be modeled as retaining the W most recent mappings it has formed, where W depends on the relevant LTD parameters, numbers of synapses, cells, etc. Furthermore, a wide range of simple learning laws would lead to a gradient of trace strength within this window.

A final question I would like to investigate concerns the generalization of the match computation sketched in Sec. 4.5. The current match computation compares (on non-episode-initial slices) two deterministic learned vectors, the vector of H-inputs that represents the model's expected input and the vector of F-inputs that is the filtered representation of the actual input (from earlier cortices). However, the output of the hippocampal loop on any given time slice could also be included in the match computation. Given the assumption that information decays out of hippocampus with some time course, addition of this signal to the match process provides a way for the recent short-term statistics of the input stream to influence the process of choosing new EC representations. This would provide a potential basis for explaining the fact that our expectations are heavily influenced by the current context (including the recent past).

4.14 Speculative Mechanism for Modulating Generality of Response

The simulation results of this chapter show some basic semantic memory properties: similarity-based generalization and categorization. However, as indicated in Sec. 1.1, the domain of semantic memory encompasses many more detailed processes that we typically describe using terms such as logical inferencing (either deductive or inductive), analogical reasoning, etc. In this section, we provide a speculative mechanism whereby modulation, on a slice-by-slice basis, of the model's ${}^{H}\theta$ parameter can vary the instantaneous generality of the representation. This is an instance of a post-computational (Brooks, 1987) mechanism for achieving category-level behavior. The sequential nature of the computation together with its interpretation as variation between general and specific information captures certain essences of the human 'stream of thought'. Three generalization modes in which the human thought process might function are listed below. These are not intended to be an exhaustive set of cognitive modes. A detailed example of how the model can move from one mode to the other by manipulation of ${}^{H}\theta$ then follows. Before describing the three modes we need the following definitions. A *time-indexed* feature is a featural occurrence at a specific temporal offset from the episode-initial time slice. A *prefix* of an episode is any subsequence of the episode that includes the episode-initial time slice. A *probe* is an episode, possibly consisting of a single time slice, which is used to prompt the network to give some output. We refer to the output that the network gives in response to a probe as the probe's *completion*. The nature of the completion will vary depending on how ${}^{H}\theta$ is varied during the read-out.

When a person is prompted by some probe, *p*, the completion may fall into one of the three following generalization modes.

a) (Mode 1) If the probe, *p*, is a prefix of some previously experienced episode, e, the person may attempt to respond with the precise remainder of *e*; that is, with its *episodic completion*. In terms of the TEMECOR models, this mode is achieved by setting ${}^{H}\theta$ to one less than the number of L2 cells active on each time slice.

b) (Mode 2) Assuming the set of previously experienced episodes are naturally clustered according to similarity, and assuming *p* has the episodic completion e, the person may attempt to recall more features (or time-indexed features) than occurred in *e* but which nevertheless are highly correlated with *e*. For example, students sometimes enter such a

mode when answering essay questions on an exam. Furthermore, when in such a mode, humans normally tend to recall the most highly correlated features first and then progressively less correlated features. In terms of the TEMECOR models, this mode is achieved by setting ${}^{H}\theta$ lower than the value yielding episodic completion.

c) (Mode 3) Again, assuming a natural clustering of the episodes, and assuming *p* has the episodic completion *e*, the person may attempt to recall only the subset of features of *e* that are most highly correlated with the set of episodes clustered with *e*. In terms of the TEMECOR models, this mode is achieved by increasing ${}^{H}\theta$. However, this implies that prior to increasing ${}^{H}\theta$, a greater-than-normal number of L2 cells has to be active in order to generate the higher ϕ values needed to exceed the increased ${}^{H}\theta$. As we narrow the search—i.e., as we raise ${}^{H}\theta$, those features that are least correlated with (i.e., have least *cue validity* for) the category of e should drop out first and the most correlated features should drop out last.

These three modes can be alternatively described as reading out a) exactly one particular set of time-indexed features corresponding to a previously learned episode, b) reading out a superset of that set of time-indexed features, and c) reading out a subset of that set of time-indexed features. The model used to illustrate the retrieval of higher-order correlational information present over the set of exemplars—specifically, modes 2 and 3 above—is a pared-down version of TEMECOR's L2. In particular, the cells comprising this model are not broken up into CMs. The single layer of the model serves as both the input and internal representation layers. The cells of the model are completely connected and spatiotemporal traces are laid down in this model according to the same Hebbian rule used in the TEMECOR models.

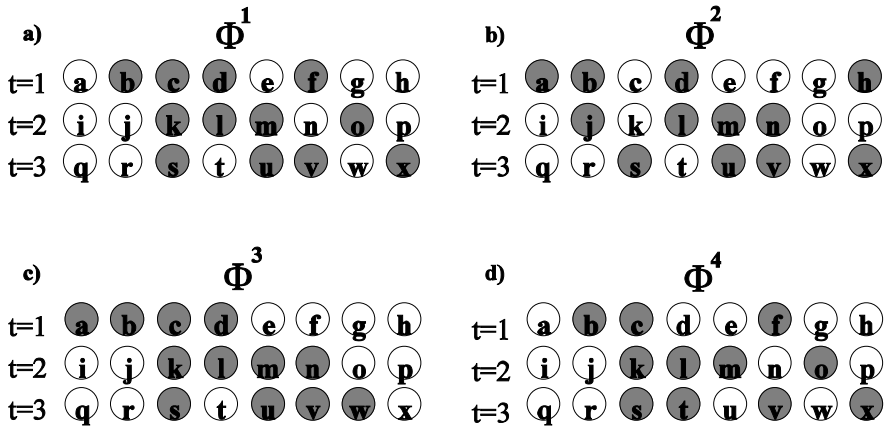

Figure 4.34: Four different episodes that are defined as instances of some single category, C_1. Note that the cells in this figure should be interpreted as L2 cells even though it is not essential for the point being made here that they exist within CMs. Also, we will not explicitly show axons and synapses in these figures however the example assumes the same simple Hebbian law used throughout the thesis so far. Each episode has three time slices as indexed. Thus, the active cells in the top row increase their weights onto the active cells in the middle row, which then increase their weights onto those in the bottom row.

Consider the set of four episodes depicted in Figure 4.34. Let these four episodes all be exemplars of a single category, C_1. These four episodes were deliberately constructed so that certain features would be more highly correlated with—i.e., more prototypical of—the category as a whole (as defined by the set of four instances) than others. Thus features l and m are always present on the second time slice of any instance of C_1, feature k is present on the second time slice of three of the instances, feature o on the second time slice of two, etc. Assuming the simple Hebbian weight increase rule described earlier, and binary-valued horizontal weights, it is the case that after presentation of all four episodes, cells l and m would each have 7 of their incoming H-wts increased, cell k would have 5 increased, cell n 6, and so forth. To avoid clutter, the actual physical connections between cells are not shown.

Figure 4.35 depicts the read-out that would result for six different settings of ${}^{H}\theta$ when prompted with the first time slice of Γ^3 (panels a-c), or with a prompt containing many intrusions from other episodes in C_1 (panels d-f). In Figure 4.35a, ${}^{H}\theta = 4$. This shows the model operating in mode 1. It performs, to the extent possible given this particular set of episodes, an episodic completion of the

probe episode, Γ^3. Note that when ${}^{H}\theta = 4$, feature x intrudes on the final time slice. In fact, there is no setting of ${}^{H}\theta$ that allows perfect recall of Γ^3 given this set of episodes.

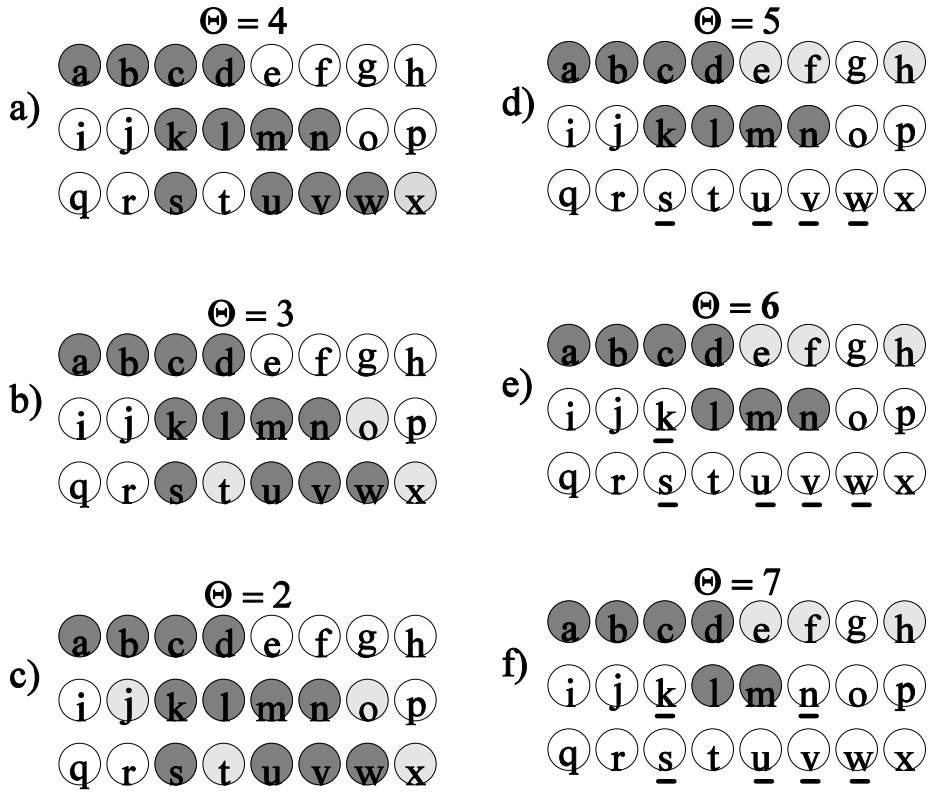

Figure 4.35: This figure shows the read-out that would result for six different settings of ${}^{H}\theta$ when prompted with the first time slice of Γ^3 (panels a-c), or with a prompt containing many intrusions from other episodes in the same category (panels d-f). Correctly activated features are darkly shaded, intruding features are lightly shaded and deleted features are underlined. The time indices are omitted in this figure but are implicitly the same as in the previous figure.

Figure 4.35b shows what happens when ${}^{H}\theta$ is lowered to three. In this case, feature o intrudes at $t = 2$ and features, t and x, intrude at $t = 3$. Figure 4.35c shows that another feature, j, which is less correlated with the category as a whole than feature o, intrudes at $t = 2$ if ${}^{H}\theta$ is lowered to two. Comparison of panels a, b and c of Figure 4.35 therefore shows that as we lower ${}^{H}\theta$, progressively less correlated features exceed ${}^{H}\theta$ and become active. This is the basic property underlying mode 2 as described in the list above. The result of processing under a reduced ${}^{H}\theta$ for several time slices is

likely to be a Δ pattern which has intrusions: we refer to such a Δ pattern as an *enriched* time slice or, insofar as any time slice can be viewed as a prompt for the next Δ pattern, an *enriched prompt*.

Figures 4.35d-f show what happens for progressively higher values of ${}^{H}\theta$ when given an enriched prompt. Specifically, we see that those features that are most highly correlated with the category, l and m, remain in the recalled trace the longest. Feature k drops out when ${}^{H}\theta = 6$ and feature n drops out when ${}^{H}\theta = 7$. Assuming the simplest approximation to a true measure of spatiotemporal correlation between a feature and the category—i.e., simply the total number of times the feature occurs across all instances of the category—then this is admittedly a violation of the desired property in that the more correlated feature, k (occurred in three episodes), drops out before the less correlated feature, n (occurred in only two episodes), as we raise ${}^{H}\theta$. However, the basic statistical trend is still present—i.e., the features, k and n, are both less correlated with the category as a whole, and they both drop out before features, l and m, which are more correlated with the category. In fact, features k and n drop out in the expected order (based on the simple correlation measure we are using) if the different enriched prompt shown in Figure 4.36, is presented. Thus, Figures 4.35d-f and 4.36 portray the model operating in generalization mode 3.

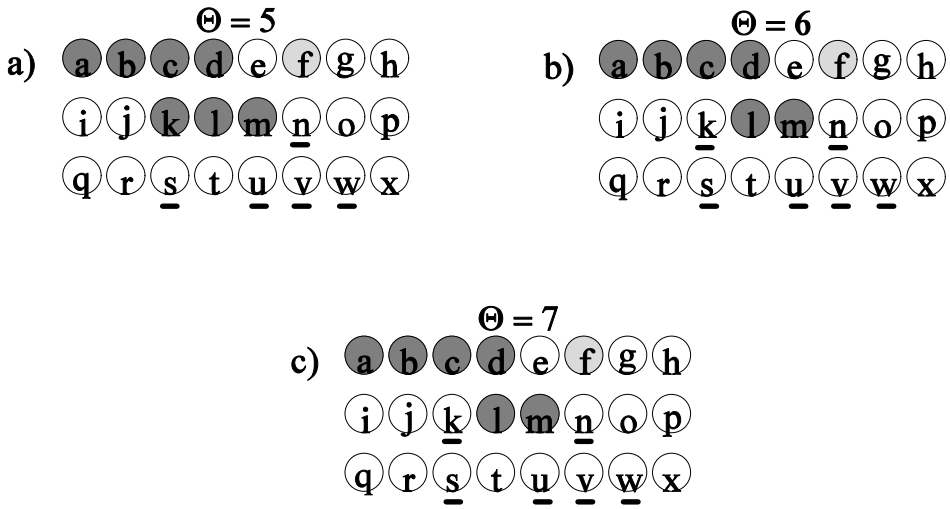

Figure 4.36: This figure shows the read-out that would result for various settings of ${}^{H}\theta$ when a different prompt (than in panels d-f of previous figure), which is also a superset of Γ_1^3, is presented

The model can be made to capture (i.e., encode) the spatiotemporal correlation inherent in the set of input episodes more accurately if a different learning dynamics is assumed. In particular, we can change the model so that (a) the weights are continuous-valued and (b) they do not go to asymptote on the first increase, but rather, subsequent weight increases to the same synapse are smaller and smaller. In this case, the weights from cells b and c onto cell k would be higher than those from b and c onto n (because cells b and c are both active immediately before cell k on three of the episodes, but before cell n only on 2). This would act to raise $\phi(k)$ and lower $\phi(n)$, relative to each other, in the examples of Figures 4.35 and 4.36, and thus make the model's behavior more likely to adhere to the true spatiotemporal correlation in the input set.

Taken together, the examples of this section suggest the following hypothesis about the neural correlates of some aspects of the human thought process. Given some initial prompt, *qua* problem statement, some swath of neural activity, corresponding to the problem solution process, will arise. ${}^{H}\theta$ may be alternately lowered and raised while this swath obtains. This modulation of ${}^{H}\theta$ partially determines this swath. During those periods in which ${}^{H}\theta$ is lowered, the swath, *qua* 'search', widens, allowing progressively less related features to become active. During those periods in which ${}^{H}\theta$ is raised, the search narrows, but crucially, as shown in Figures 4.35 and 4.36, those features which are most likely to remain active are those most highly correlated with the category represented by the set of cells that were active immediately prior to the rise in ${}^{H}\theta$. Because this example involved a very pared-down version of the model, much future work is necessary in order to fully develop this mechanism.

Chapter 5. Conclusion

An unsupervised, distributed, two-layer, associative network model of storage, retrieval and recognition of binary spatiotemporal patterns has been described. The model was developed in two stages. The first version, TEMECOR-I, exhibits several essential properties of episodic memory: very high capacity, single-trial learning, permanence (i.e., stability) of traces, and the ability to store highly overlapped spatiotemporal patterns, including complex state sequences (CSSs). However, due to the fact that it fails to have continuity in either the F-projection, which is a spatial mapping from inputs to internal representations (IRs), or the H-projection, which maps IRs to successor IRs, it provides no basis for similarity-based recognition and categorization, which form the basis of semantic memory. The second version, TEMECOR-II, does achieve continuity in both the F- and H-projections and thus exhibits similarity-based recognition and categorization for spatiotemporal patterns. TEMECOR-II retains the episodic memory properties of its predecessor (although its capacity is significantly lower) and thus constitutes a single model that exhibits properties of three major phenomenal domains of memory: episodic, semantic and spatiotemporal (i.e., sequence) memory. Another important feature of the model is that it is monolithic. That is, it is organized into two homogeneous fields of cells, L1 and L2. In fact, it is largely a local circuit model, with the competitive module (CM) being the basic local circuit.

The theory has been inspired by some of the more general architectural and dynamical properties of the cortex of the mammalian brain. The model's layer two (L2) is considered to be analogous to the brain's entorhinal cortex (EC) and the model's L1 is considered analogous to the various higher-level association cortices that feed into EC. The model makes very simple assumptions about the fundamental constituents of the brain, neurons and synapses. Specifically, it assumes binary activation values for neurons and binary-valued synapses. Simple Hebbian learning of an all-or-none variety is used. The model utilizes the reciprocal innervation that is found throughout the cortex, but does not require strict, symmetrical reciprocal connectivity.

In order to achieve very high capacities the model needs to assume a connectivity rate in the H-projection that is biologically unrealistic. A speculative solution for this problem, involving the hippocampus, was outlined in Sec. 4.13. This hypothesis is that the hippocampus effectively

increases the connectivity rate of EC by providing multi-synaptic pathways from any point in EC to any other point in EC.

TEMECOR-II is based on three fundamental computational principles: a) a sparse, distributed representation scheme; b) computation of the degree of match, *G*, between the expected and actual inputs at *t*; and c) addition of a graded amount of noise, Λ, inversely dependent on *G*, into the process of choosing winning cells in each of the CMs. This generally leads to reactivation of old traces (i.e., greater pattern completion) in proportion to the familiarity of inputs, and establishment of new traces (i.e., greater pattern separation) in proportion to the novelty of inputs. This is tantamount to the property of continuity in the input-to-IR map and it endows the model with recognition and categorization capability as demonstrated in the simulations of Sec. 4.10.5.

The model has several other properties desirable from a psychological point of view. It has the property that it can read out the remainder of a previously encountered sequence if given a prompt that originally occurred in a mid-sequence position. For example, if the model has previously learned the state sequence, [A B C D], then if it is prompted with B, it will complete with [C D]. Furthermore, this generalizes to the case in which it is given a spatiotemporal prompt so that if prompted with [B C], it will complete with D. Any realistic model of the storage and processing of sequential information in humans must have this property. In contrast, models based on recurrent backpropagation are unlikely to have this capability. This is because presentation of a mid-sequence prompt to an RBP-based model would require the presence of the correct pattern over the context (i.e., state) units in order for the correct completion to obtain, however that context pattern can only come to exist by the model passing through all the prior states of the sequence involved.

The model is also capable of representing multiple competing hypotheses and deferring judgment until subsequent disambiguating information enters the system. For example, if the model has previously learned the two sequences, Γ^1 = [A B C D E] and Γ^2 = [F B G H I], then if it is prompted with state B, two internal representations of state B, Δ^1_B and Δ^2_B, will become co-active. If state C presents on the next time step then the model will lock into the memory trace for Γ^1. If state *G* presents, it will lock into the trace for Γ^2. The model automatically detects and handles such ambiguities. Furthermore, this capability generalizes to the case where the ambiguity persists for multiple time steps. Thus, if it has learned Γ^1 = [A B C D E] and Γ^2 = [F B C H I], then if presented with [B C H], it would entertain competing hypotheses for the first two time slices of the prompt and then lock into and read out the remainder of Γ^2.

An additional feature of TEMECOR-II is that it has two major operational modes that are intended to correspond to two major modes of human cognition. Conscious processing is sometimes highly dependent on immediate feedback, from one moment in time to the next, from the environment. The activity of reading has this character. It is a sequential process in which new input enters the system on successive moments. We have labeled this mode of conscious processing as interactive mode. At other times, conscious processing is relatively insensitive to external inputs; e.g., reminiscence or daydreaming. We have denoted this as solipsistic mode. An additional feature of TEMECOR-II is that it has two major operational modes that are intended to correspond to the interactive and solipsistic modes of human cognition. At this point, both operational modes are developed. However, a more comprehensive model including the many other factors that control the movement between modes, e.g., mood, etc., is a matter for future research.

The proposed model was originally developed to account for episodic recall of spatiotemporal patterns. The program of research has now led to a more comprehensive model that exhibits some of the basic properties of semantic memory as well. Many avenues of exploration are now possible. The last two sections of Ch. 4, regarding inclusion of a hippocampal analog to the model and a potential means for controlling the generality of the information contained in the unfolding memory trace, appear to be particularly promising.

References

Albus, James S. (1971) "A theory of cerebellar functions" *Mathematical Biosciences*, **10**(1/2), 25-61

Alvarez, R. & Squire, L. R. (1994**)** "Memory consolidation and the medial temporal lobe: a simple network model" *Proceedings of the National Academy of Sciences, **91***, 7041-7045.

Amit, D. J. (1988**)** *Modeling Brain Function: The World of Attractor Neural Networks*. Cambridge University Press, Cambridge, UK.

Anderson, J. A. (1983**)** "Cognitive and Psychological Computation with Neural Models" *IEEE Trans. on Systems, Man and Cybernetics*, ***SMC-13***, 799-815.

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

Anderson, J. R. & Bower, G. H. (1973**)** *Human Associative Memory*. Winston, Washington, DC.

Ans, B., Coiton, Y., Gilhodes, J-C. & Velay, J-L. (1994**)** "A Neural Network Model for Temporal Sequence Learning and Motor Programming" *Neural Networks, **7***, 1461-1476.

Bradski, G., Carpenter, G. A. & Grossberg, S. (1992**)** "Working Memory Networks for Learning Temporal Order with Application to Three-Dimensional Visual Object Recognition" *Neural Computation, **4***, 270-286.

Brooks, L. R. (1987**)** "Decentralized Control of Categorization: the role of prior processing episodes" In *Concepts and Conceptual Development*. Ch. 6, 141-174. Cambridge University Press. Papers from the Emory Symposia in Cognition 1 (Oct. 1984).

Brooks, L. R. (1978**)** "Nonanalytic Concept Formation and Memory for Instances" In *Cognition and Categorization*. Rosch, E. & Lloyd, B.B. (Eds.) Ch. 7, 169-215. Lawrence Erlbaum Associates. NJ.

Bruce, C., Desimone, R. & Gross, C. G. (1981**)** "Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque" *Journal of Neurophysiology, **46***, 369-384.

Carpenter, G. & Grossberg, S. (1993**)** "Normal and amnesic learning, recognition and memory by a neural model of cortico-hippocampal interactions" *Trends in Neuroscience, **16***, 131-137.

Carpenter, G. & Grossberg, S. (1987**)** "Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine" *Computer Vision, Graphics and Image Processing, **37***, 54-115.

Cleeremans, A. (1993**)** *Mechanisms of Implicit Learning: Connectionist Models of Sequence Processing*. A Bradford Book, The MIT Press, Cambridge, MA.

Cleeremans, A. & McClelland, J. L. (1990**)** "Learning the Structure of Event Sequences" In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum Associates, NJ. 709-716.

Cohen, G. (1989**)** *Memory in the Real World*. Lawrence Erlbaum Associates, NJ.

Cohen, M. & Grossberg, S. (1987**)** "Masking Fields: A Massively Parallel Neural Architecture for Learning, Recognizing, and Predicting Multiple Groupings of Patterned Data" *Applied Optics*, ***26****(10)*, 1866-1891.

Collins, A. M. & Loftus, E. F. (1975**)** "A spreading activation theory of semantic processing" *Psychological Review*, ***82***, 407-428.

Collins, A. M. & Quillian, M. R. (1969**)** "Retrieval time from semantic memory" *Journal of Verbal Learning and Verbal Behavior*, ***8***, 240-247.

Coultrip, R. L. & Granger, R. H. (1994**)** "Sparse Random Networks With LTP Learning Rules Approximate Bayes Classifiers via Parzen's Method" *Neural Networks*, ***7***, 463-476.

Eccles, J. C. (1981**)** "The Modular Operation of the Cerebral Neocortex Considered as the Material Basis of Mental Events" *Neuroscience*, ***6***, 1839-1856.

Elman, J. L. (1991**)** "Distributed Representations, Simple Recurrent Networks, and Grammatical Structure" *Machine Learning*, ***7***, 91-122 .

Elman, J. L. (1990**)** "Finding Structure in Time" *Cognitive Science*, ***14***, 179-212.

Engel, A. K., Konig, P., Kreiter, A. K., Schillen, T. B. & Singer, W. (1992**)** "Temporal Coding in the Visual Cortex: new vistas on integration in the nervous system" *Trends in Neuroscience*, *15*.

French, R. M. (1994**)** "Dynamically constraining connectionist networks to produce distributed, orthogonal representations to reduce catastrophic interference" In *Proceedings of the 16th Annual Conference of the Cognitive Science Society*, Lawrence Erlbaum Associates, NJ. 335-340.

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

Fujita, I., Tanaka, K., Ito, M. & Cheng, K. (1992**)** "Columns for visual features of objects in monkey inferotemporal cortex" *Nature*, ***360***, 343-346.

Gabriel, M., Sparenborg, S. P. & Stolar, N. (1986**)** "An Executive Function of the Hippocampus: Pathway Selection for Thalamic Neuronal Significance Code" In *The Hippocampus **4***, Isaacson, R.L. & Pribram, K.H. (Eds.) Ch. 1, 1-37.

Gross, C. G. & Rocha-Miranda, C. E. and Bender, D.B. (1972**)** "Visual properties of neurons in the inferotemporal cortex of the monkey" *Journal of Neurophysiology, **35***, 96-111.

Grossberg, S. (1986**)** "The Adaptive Self-Organization of Serial Order in Behavior: Speech, Language, and Motor Control" In *Pattern Recognition by Humans and Machines: Speech Perception, 1*, Academic Press. Ch. 6, 187-293.

Grossberg, S. (1982**)** "Processing of Expected and Unexpected Events during Conditioning and Attention: A Psychophysiological Theory" *Psychological Review, **89***, 529-572.

Grossberg, S. (1980**)** "How does the Brain Build a Cognitive Code" *Psychological Review, **87***, 1-51.

Grossberg, S. (1978**)** "A Theory of Human Memory: Self-Organization and Performance of Sensory-Motor Codes, Maps, and Plans" *Progress in Theoretical Biology, **5***, 233-374.

Grossberg, S. (1976**)** "Adaptive Pattern Classification and Universal Recoding II: Feedback, Expectation, Olfaction, Illusions" *Biological Cybernetics, **23***, 187-202.

Grossberg, S. (1973**)** "Contour Enhancement, Short Term Memory, and Constancies in Reverberating Neural Networks" *Studies in Applied Mathematics (L II)*, 213-257.

Hasselmo, M.E., Schnell, E. & Berke, J. & Barkai, E. (1995**)** "A model of hippocampus combining rapid self-organization and associative memory function" In *Advances in Neural Information Processing Systems 7*, Touretzkey, D., Mozer, M. & Leen, T. (Eds.) The MIT Press, Cambridge, MA.

Hasselmo, M. E. (1995**)** "Neuromodulation and cortical function: modeling the physiological basis of behavior" *Behavioural Brain Research, **67***, 1-27.

Hasselmo, M. E. (1994**)** "Runaway Synaptic Modification in Models of Cortex: Implications for Alzheimer's Disease" *Neural Networks, **7***, 13-40.

Hasselmo, M. E. (1993**)** "Acetylcholine and Learning in a Cortical Memory" *Neural Computation, **5***, 32-44.

Hasselmo, M. E., Anderson, B. P. & Bower, J. M. (1992**)** "Cholinergic Modulation of Cortical Associative Memory Function" *Journal of Neurophysiology, **67*** , 1230-1246.

Hasselmo, M. E., Anderson, B. P. & Bower, J. M. (1991**)** "Cholinergic Modulation may Enhance Cortical Associative Memory Function" *Journal of Neurophysiology*, 46-52.

Hasselmo, M. E., Rolls, E. T. & 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" *Exp. Brain Research, **32***, 203-218.

Hasselmo, M. E. & Stern, C. E. (1995**)** "Linking LTP to network function: A simulation of episodic memory in the hippocampal system" In *Long-term potentiation, Vol. 3*. Baudry, M. & Davis, J. (Eds.) MIT Press, Cambridge, MA.

Hebb, D. O. (1949**)** *The Organization of Behavior*. Wiley, NY.

Hinton, G. E., McClelland, J. L. & Rumelhart, D. E. (1986**)** "Distributed Representations" In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* Vol. *1*, McClelland, J., Rumelhart, D. & The PDP Research Group (Eds.) The MIT Press. 77-109.

Hintzman, Douglas L. (1986**)** "'Schema Abstraction' in a Multiple-Trace Memory Model" *Psychological Review, **93***, 411-428.

Hochreiter, S. & Schmidhuber, J. (1995**)** "Long Short Term Memory" Tech. Rep. FKI-207-95, FKI, Germany.

Homa, D., Cross, J., Cornell, D., Goldman, D. & Schwartz, S. (1973**)** "Prototype abstraction and classification of new instances as a function of number of instances defining the prototype" *Journal of Experimental Psychology, **101***, 116-122.

Hopfield, J. J. (1984**)** "Neurons with Graded Responses have collective computational properties like those of two-state neurons" *Proceedings of the National Academy of Science, **81***, 3088-3092.

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

Hopfield, J. J. & Tank, D. W. (1989**)** "Neural Architecture and Biophysics for Sequence Recognition" *In Neural Models of Plasticity: Experimental and Theoretical Approaches*. Byrne, J. H. & Berry, W. O. (Eds.) Ch. 17, 363-377.

Hubel, D. H. & Wiesel, T. N. (1968**)** "Receptive fields and functional architecture of the monkey striate cortex" *Journal of Physiology, **195***, 215-243.

Jordan, M. I. (1986**)** "Serial Order" Tech. Rep. 8604, Institute for Cognitive Science, University of California, San Diego, CA.

Kelso, S. R., Ganong, A. H. & Brown, T. H. (1986**)** "Hebbian Synapses in the Hippocampus." *Proc. of the National Academy of Science, **83***, 5326-5330.

Kobatake, E. & Tanaka, K. (1994**)** "Neuronal Selectivities to Complex Object Features in the Ventral Visual Pathway of the Macaque Cerebral Cortex" *Journal of Neurophysiology*, ***71***(3), 856-867.

Kohonen, T. (1989**)** *Self-Organization and Associative Memory* (3rd Ed.) Springer Series in Information Sciences. Springer-Verlag, Berlin, Germany.

Kohonen, T. (1972**)** "Correlation Matrix Memories", *IEEE Trans. on Computers*, ***C-21***, 353-359.

Kortge, C. A. (1990**)** "Episodic Memory in Connectionist Networks" *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, Lawrence Erlbaum Associates, NJ. 764-771.

Kruschke, J. K. (1992**)** "ALCOVE: An Exemplar-Based Connectionist Model of Category Learning" *Psychological Review*, ***99****(1)*, 22-44.

Lashley, K. S. (1951**)** "The problem of serial order in behavior" In *Cerebral Mechanisms of Behavior*. Jeffress, L.A. (Ed.) Wiley, NY. 112-136.

Levy, W. B. (1985) "Associative changes in the synapse: LTP in the hippocampus" In *Synaptic Modification, Neuron Selectivity, and Nervous System Organization*. Levy, W. B., Anderson, J. & Lehmkuhle, S. (Eds.) Lawrence Erlbaum Associates, NJ. 105-121.

Levy, W. B. (1989**)** "A Computational Approach to Hippocampal Function" In *Computational Models of Learning in Simple Neural Systems*. Vol. 23. Hawkins, R. D. & Bower, G. H. (Eds.) Academic Press. 243-305.

Levy, W. B., Colbert, C. M. & Desmond, N. L. (1990**)** "Elemental Adaptive Processes of Neurons and Synapses: A Statistical/Computational Perspective" In *Neuroscience and Connectionist Theory*. Gluck, M. & Rumelhart, D. (Eds.) Lawrence Erlbaum Associates, NJ. Ch. 5, 187-235.

Levy, W. B. & Desmond, N. (1985**)** "The Rules of Elemental Synaptic Plasticity" In *Synaptic Modification, Neuron Selectivity, and Nervous System Organization*. Levy, W. B., Anderson, J. & Lehmkuhle, S. (Eds.) Lawrence Erlbaum Associates, NJ. Ch. 6.

Levy, W. B., Wu, X. & Baxter, R. A. (1995**)** "Unification of hippocampal function via computational/encoding considerations" *International Journal of Neural Systems*, ***6***, 71-80.

Loftus, E. F. (1977**)** "Shifting Human Color Memory" *Memory and Cognition*, *6*, 696-699.

Luria, A. R. (1968**)** *The Mind of a Mnemonist* (translated by Lynn Solotaroff). Basic Books.

Lynch, G. (1986**)** *Synapses, Circuits, and the Beginnings of Memory*. The MIT Press, Cambridge, MA.

Lynch, G. & Granger, R. (1994**)** "Variations in Synaptic Plasticity and Types of Memory in Corticohippocampal Networks" In *Memory Systems 1994*. Schacter, D. S. & Tulving, E. (Eds.) A Bradford Book, The MIT Press. Cambridge, MA. Ch. 3, 65-86.

Marr, D. (1969**)** "A Theory of Cerebellar Cortex" *Journal of Physiology*, ***202***, 437-470.

Maunsell, J. H. R. & Van Essen, D. C. (1983**)** "The connections of the middle temporal visual area (MT) and their relation to a cortical hierarchy in the monkey" *Journal of Neuroscience*, ***3***, 2563-2586.

McClelland, J. L. (1986**)** "Resource Requirements of Standard and Programmable Nets" In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. *1*, McClelland, J., Rumelhart, D. & The PDP Research Group (Eds.) The MIT Press. Ch. 12, 460-487.

McClelland, J. L. & Rumelhart, D. E. (1985**)** "Distributed Memory and the Representation of General and Specific Information" *Journal of Experimental Psychology: General*, ***114***, 159-188.

McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. (1994**)** "Why there are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the successes and Failures of Connectionist Models of Learning and Memory" Tech. Rep. PDP.CNS.94.1, Parallel and Distributed Processing and Cognitive Neuroscience, Carnegie-Mellon University, Pittsburgh, PA.

McCloskey, M. & Cohen, N. J. (1989**)** "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem", In *The Psychology of Learning and Memory Vol. 24*. Bower, G. H. (Ed.) Academic Press. 109-165.

McRae, K. & Hetherington, P. A. (1993**)** "Catastrophic Interference is eliminated in pretrained networks" In *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, Lawrence Erlbaum Associates, NJ. 723-728.

Medin, D. L. & Schaffer, M. M. (1978**)** "Context Theory of Classification Learning" *Psychological Review*, ***85****(3)*, 207-238.

Miller, R. (1991**)** *Cortico-Hippocampal Interplay and the Representation of Contexts in the Brain*. Springer-Verlag, Heidelberg, Germany.

Minai, A. A., Barrows, G. L. & Levy, W. B. (1994**)** "Disambiguation of pattern sequences with recurrent networks" In *Proceedings of the INNS World Congress on Neural Networks Vol. IV*, Lawrence Erlbaum Associates, NJ. 178-181.

Minai, A. A. & Levy, W. B. (1993**)** "Sequence Learning in a single-trial" In *Proceedings of the INNS World Congress on Neural Networks Vol. II*, Lawrence Erlbaum Associates, NJ. 505-508.

Moll, M. & Miikkulainen, R. (1995**)** "Convergence-Zone Episodic Memory: Analysis and Simulations" Tech. Rep. AI95-227, The University of Texas at Austin, Dept. of Computer Science.

Moll, M., Miikkulainen, R. & Abbey, J. (1993**)** "The Capacity of Convergence-Zone Episodic Memory" Tech. Rep. AI93-210, The University of Texas at Austin, Dept. of Computer Science.

Mountcastle, V. B. (1978**)** "An organizing principle for cerebral function: the unit module and the distributed system" In *The Mindful Brain*. Schmitt, F. O. (Ed.) The MIT Press, Cambridge, MA. 7-50.

Muller, R. U., Kubie, J. L., Bostock, E. M., Taube, J. S. & Quirk, G. L. (1991**)** "Spatial Firing Correlates of Neurons in the Hippocampal Formation of Freely Moving Rats" In *Brain and Space*. Paillard, J. (Ed.) Ch. 17. Oxford University Press, Oxford.

Murre, J. (1995**)** "A Model of Amnesia" *Psychological Review (submitted).*

Murre, J. (1992**)** *Learning and Categorization in Modular Neural Networks*. Lawrence Erlbaum Associates, NJ.

Nadal, J-P. & Toulouse, G. (1990**)** "Information storage in sparsely coded memory nets" *Network*, *1*, 61-74.

Neisser, U. (1982**)** "Memorists" In *Memory Observed*. Neisser, U. (Ed.) Ch. 37. W. H. Freeman and Company, 377-381.

O'Keefe, J. & Conway, D. H. (1978**)** "Hippocampal place units in the freely-moving rat: Why they fire and where they fire" *Experimental Brain Research*, ***31***, 573-590.

O'Keefe, J. & Nadel, L. (1978**)** *The Hippocampus as a Cognitive Map*. Clarendon Press, Oxford, England.

Oldfield, R. C. (1963**)** "Individual Vocabulary and Semantic Currency", *British Journal of Social and Clinical Psychology*, ***2***, 122-130.

O'Reilly, R. C. & McClelland, J. L. (1994**)** "Hippocampal Conjunctive Encoding, Storage and Recall: Avoiding a Tradeoff" Tech. Rep. PDP.CNS.94.4, Parallel and Distributed Processing and Cognitive Neuroscience, Carnegie-Mellon University, Pittsburgh, PA.

Palm, G. (1980**)** "On Associative Memory" *Biological Cybernetics*, ***36***, 19-31.

Perrett, D. I., Rolls, E. T. & Caan, W. (1982**)** "Visual neurones responsive to faces in the monkey temporal cortex" *Experimental Brain Research*, ***47***, 329-342.

Posner, M. I. & Keele, S. W. (1970**)** "Retention of Abstract Ideas" *Journal of Experimental Psychology*, ***83***, 304-308.

Posner, M. I. & Keele, S. W. (1968**)** "On the genesis of abstract ideas" *Journal of Experimental Psychology*, ***77***, 353-363.

Quillian, M. R. (1966**)** *Semantic Memory*. Ph.d. Thesis, Carnegie Institute of Technology.

Quillian, M. R. (1968**)** "Semantic Memory" In *Semantic Information Processing*. Minsky, M. (Ed.) The MIT Press.

Reber, A. S. (1976**)** "Implicit Learning of Synthetic Languages: The Role of Instructional Set" *Journal of Experimental Psychology: Human Learning and Memory*, ***2(1)***, 88-94.

Reber, A. S. (1967**)** "Implicit learning of artificial grammars" *Journal of Verbal Learning and Verbal Behavior*, ***5***, 855-863.

Reiss, M. & Taylor, J. G. (1991**)** "Storing Temporal Sequences" *Neural Networks*, ***4***, 773-787.

Ribot, T. (1882**)** *The Diseases of Memory*. Appleton, New York.

Rinkus, G. (1995**)** "TEMECOR: An Associative, Spatio-temporal Pattern Memory for Complex State Sequences" In *Proceedings of the 1995 World Congress on Neural Networks, Vol. I,* Lawrence Erlbaum Associates and The INNS Press. 442-448.

Rinkus, G. (1993**)** "Context-sensitive Spatio-temporal Memory" In *Proceedings of the 1993 World Congress on Neural Networks Vol. II*, Lawrence Erlbaum Associates and The INNS Press. 344-347.

Rockel, A. J., Hiorns, R. W. & Powell, T. P. S. (1980) "The basic uniformity in structure of the neocortex" *Brain*, ***103***, 221-244.

Rockland, K. S. & Pandya, D. N. (1979**)** "Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey" *Brain Research*, ***179***, 3-20.

Rolls, E. T. (1990**)** "Functions of Neuronal Networks in the Hippocampus and of Backprojections in the Cerebral Cortex in Memory" In *Neural Models of Plasticity: Experimental and Theoretical Approaches*. Ch. 13. Byrne, J. H. & Berry, W. O. (Eds.) Academic Press. 184-210.

Rolls, E. T. (1990**)** "Functions of Neuronal Networks in the Hippocampus and Neocortex in Memory" In *Brain Organization and Memory: Cells, Systems, and Circuits*. Ch. 9, McGaugh, J., Weinberger, N. & Lynch, G. (Eds.) Oxford University Press. 240-265.

Rubin, D. C. & Kozin, M. (1984**)** "Vivid Memories" *Cognition*, ***16***, 81-95.

Rumelhart, D. E. & McClelland, J. L. (1986**)** "On Learning the Past Tenses of English Verbs" In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 2, McClelland, J., Rumelhart, D. & The PDP Research Group (Eds.) The MIT Press. 216-271.

Rumelhart, D. E. & Norman, D. A. (1978**)** "Accretion, Tuning, and Restructuring: Three Modes of Learning" In *Semantic Factors in Cognition*. Cotton, J. & Klatzky, R. (Eds.) Lawrence Erlbaum Associates, NJ. 37-53.

Schank, R. C. (1982**)** *Dynamic Memory*. Cambridge University Press, Cambridge, England.

Servan-Schreiber, D., Cleeremans, A. & McClelland, J. L. (1991**)** "Graded State Machines: The Representation of Temporal Contingencies in Simple Recurrent Networks", *Machine Learning*, ***7***, 161-193.

Shapiro, M. L. & Olton, D. S. (1994**)** "Hippocampal Function and Interference" In *Memory Systems 1994*. Schacter, D. S. & Tulving, E. (Eds.) A Bradford Book, The MIT Press. Cambridge, MA. Ch. 4, 87-117.

Shaw, G. L., Harth, E. & Scheibel, A. B. (1982**)** "Cooperativity in Brain Function: Assemblies of Approximately 30 Neurons" *Experimental Neurobiology*, ***77***, 324-358.

Smith, A. W. & Zipser, D. (1989**)** "Learning Sequential Structure with the Real-Time Recurrent Learning Algorithm" *International Journal of Neural Systems*, ***1***, 125-131.

Smith, E. E. & Medin, D. L. (1981**)** *Categories and Concepts*. Harvard University Press, Cambridge, MA.

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

Squire, L. R., Cohen, N. J. & Nadel, L. (1984**)** "The Medial Temporal Region and Memory Consolidation: A New Hypothesis" In *Memory Consolidation: Psychobiology of Cognition*. Weingartner, H. & Parker, E. S. (Eds.) Lawrence Erlbaum Associates, NJ. Ch. 8, 185-210.

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

Szentagothai, J. (1975**)** "The "module-concept" in cerebral cortex architecture" *Brain Research*, ***95***, 475-496.

Tanaka, K. (1993**)** "Neuronal Mechanisms of Object Recognition" *Science*, ***262***, 685-688.

Tulving, E. (1972**)** "Episodic and Semantic Memory" In *Organization of Memory*. Tulving, E. & Donaldson, W. (Eds.) Academic Press, New York.

Van Gelder, Tim (1990**)** "Compositionality: A Connectionist Variation on a Classical Theme" *Cognitive Science*, ***14***, 355-384.

Van Hoesen, G. W. & Pandya, D.N. (1975**)** "Some connections of the entorhinal (area 28) and perirhinal (area 35) cortices of the rhesus monkey. 1. Temporal lobe afferents" *Brain Research*, ***95***, 1-24.

Vogh, J. (1993**)** "Sequential Memory with ART: A Self-Organizing Neural Network Capable of Learning Sequences of Patterns" Tech. Rep. CAS/CNS-TR-93-033, Dept. of Cognitive and Neural Systems, Boston University, Boston, MA.

Vokey, J. R. & Brooks, L. R. (1992**)** "Salience of Item Knowledge in Learning Artificial Grammars" *Journal of Experimental Psychology: Learning, Memory and Cognition*, ***28***, 328-344.

Waibel, A. (1989) "Consonant and Phoneme Recognition by Modular Constructions of Large Phonemic Time-Delay Neural Networks" In *Advances in Neural Information Processing Systems, I*. Touretzky, D. S. (Ed.) Morgan Kaufmann, San Mateo, CA. 215-223.

Waibel, A., Sawai, H. & Shikano, K. (1989**)** "Consonant and phoneme recognition by modular constructions of large phonemic time-delay neural networks" In *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. & Lang, K. L. (1989**)** "Phoneme Recognition Using Time-Delay Neural Networks" In *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE.

Whittlesea, B. W. A. (1989**)** "Selective attention, variable processing, and distributed representation: preserving particular experiences of general structures" In *Parallel Distributed Processing: Implications for Psychology and Neurobiology*. Morris, R. (Ed.) Oxford Science Publications. Ch. 5, 76-101.

Whittlesea, B. W. A. (1987**)** "Preservation of Specific Experiences in the representation of general knowledge" *Journal of Experimental Psychology: Learning, Memory and Cognition*, ***13***, 3-17.

Whittlesea, B. W. A. & Dorken, M. D. (1993**)** "Incidentally, Things in General Are Particularly Determined: An Episodic-Processing Account of Implicit Learning" *Journal of Experimental Psychology: General*, ***122***, 227-248.

Wickelgren, W. A. (1969a**)** "Coding, Retrieval, and Dynamics of Multitrace Associative Memory" In *Cognition in Learning and Memory*. Gregg, L. W. (Ed.), John Wiley and Sons. Ch. 2, 19-50.

Wickelgren, W. A. (1969b**)** "Context-Sensitive Coding, Associative Memory, and Serial Order in (Speech) Behavior" *Psychological Review*, ***76(1)***, 1-15.

Wigstrom, H., Gustaffson, B., Huang, Y-Y. & Abraham, W. C. (1986**)** "Hippocampal longterm potentiation is induced by pairing single afferent volleys with intracellularly injected depolarizing currents" *Acta Physiologica Scandinavica*, ***126***, 317-319.

Williams, J. & Zipser, D. (1989**)** "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks" *Neural Computation*, ***1***, 270-280.

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

Wilson, F. A.W. & Rolls, E. T. (1990**)** "Learning and memory are reflected in the responses of reinforcement-related neurons in the primate basal forebrain" *Journal of Neuroscience*, ***10***, 1254-1267.

Wilson, M. A. & McNaughton, B. L. (1994**)** "Reactivation of Hippocampal Ensemble Memories During Sleep" *Science*, ***265***, 676-679.

Wilson, M. A. & McNaughton, B. L. (1993**)** "Dynamics of the Hippocampal Ensemble Code for Space" *Science*, ***261***, 1055-1058.

Wu, X. & Levy, W. B. (1995**)** "Controlling Performance by Controlling Activity Levels in a Model of Hippocampal Region CA3. II. Overcoming the effect of Noise by Adjusting Network Excitability Parameters" In *Proceedings of the 1995 World Congress on Neural Networks*, *1*, Lawrence Erlbaum Associates, NJ. 577-581.