Integration and Shielding of Regular and Irregular Items in MLPs

Brett Gray and **Janet Wiles**
Depts of Computer Science and Psychology
University of Queensland, QLD 4072 Australia
brettg@cs.uq.oz.au janetw@cs.uq.oz.au

Abstract

Multi-layer perceptrons (MLPs) can learn both regular and irregular items given sufficient interleaved training, but not from sequential presentation of items. McClelland, McNaughton and O'Reilly (1994) addressed this problem in their proposal that the hippocampus and neocortex (H/NC) form a two component memory system in which the hippocampus interleaves training of items to the neocortex so that it can develop structure without interference of later items on earlier ones. We have been studying such an interleaving system under the constraint of limiting the capacity of the training batch (analogous to a finite limit on the hippocampus). In previous simulations (Gray & Wiles, 1996) we demonstrated that a quasi-regular learning task trained with a recency rehearsal scheme did not suffer interference to a catastrophic level, but did suffer interference on irregular and similar regular items. The current study introduces a new rehearsal scheme in which items are retained in a finite training batch based on how well the MLP has learned them: Error rehearsal enabled the MLP to learn (1) a high proportion of the domain, (2) retention of both regular and irregular items from the initial training batch and (3) partial shielding of both regular and irregular items from later interference. The results demonstrate that although finite training batches can pose a problem for MLPs, an error rehearsal scheme can reduce interference on both regular and irregular items, even when they are no longer in the current training batch. Implications for the role of the hippocampus in interleaving items for the neocortex are discussed.

Introduction

Memory systems have several functionally different learning requirements: the primary one is the ability to store events or items after only one presentation (called one-shot learning or memorisation); a second is a longer term integrative function that allows a memory system to develop a structure over the events experienced, and hence generalise to novel events. When modelling these functions independently, artificial neural network researchers have typically used single layer networks (s.a. the Matrix Model, Humphreys, Bain & Pike, 1989, or k-Winner Take All, O'Reilly & McClelland, 1994) to model one-shot learning, and multi-layer perceptrons (*MLPs*) to model integrative components (Rumelhart, Hinton & Williams, 1995).

A central problem in modelling human and animal memory is whether the two functions can be produced by a unitary memory system or whether two components are required (Humphreys, Bain & Pike, 1989; McClelland, McNaughton & O'Reilly, 1994; O'Reilly & McClelland, 1994): Since one-shot learning systems need to unambiguously store single events at the time at which they occur, they cannot take into account information that may be distributed across several events. Attempts to use one-shot learning systems as integrative systems result in severe restrictions on the types of generalisation that can occur (e.g., linear combinations in the Matrix Model; category membership in the k-Winner Take All). By contrast, integrative systems such as MLPs can utilise higher-order statistics of events distributed over time, but are not suitable for unambiguously storing sequentially presented events. MLPs are traditionally trained by interleaving presentations of all items and attempts to use MLPs as one-shot memory systems (training without interleaving presentations of items) display spectacular failures, called "catastrophic interference" (*CI*, McCloskey & Cohen, 1989; Ratcliff, 1990). In CI, events presented at an early stage in training are completely lost during training on later items.

A key question for two-component models is the relationship between the one-shot and integrative components. Models that seek to combine both these functions have rarely been explored at a purely computational level, but have been studied by cognitive neuroscientists with respect to the hippocampus/neocortex (*H/NC*) memory system (McClelland, McNaughton & O'Reilly, 1994), and the necessity for two component models has been debated by psychologists studying human memory (see Dennis, 1994, for a history of memory and learning research) and exception/regular learning (e.g., the Dual Route model, Bakker, 1995; Coltheart *et. al.*, 1993). Our analysis of the computational questions in this project derives from the H/NC research by McClelland and colleagues (although we see many parallels in related areas of psychological research).

McClelland and colleagues (McClelland, McNaughton & O'Reilly, 1994; O'Reilly & McClelland, 1994) have proposed that events are initially stored in the hippocampus (a one-shot memory which they model as a k-Winner Take All network, O'Reilly & McClelland, 1994), which then interleaves presentation of all its stored memories to the neocortex (a structured integrative memory system analogous to an MLP). Such internal interleaving enables multiple presentation of items to the integrative component (thus providing a mechanism to mitigate interference and build structure gradually), while requiring events to be presented to the entire "two-component" system only once.

Our goal in this project is to investigate computational aspects of the interleaving process between the two functional components of memory systems such as McClelland, Mc-

Naughton and O'Reilly's H/NC model under the constraint of a finite limit on the training batch sizes, analogous to a finite limit on hippocampus (Treves & Rolls, 1994). Following McClelland, McNaughton and O'Reilly's analogy, we used an MLP as an integrative component, and a buffer as the one-shot component.

The focus of this study was the interleaving process (called a rehearsal scheme) and its effect on CI. Several rehearsal schemes have been reported in the neural network literature. Results either showed no significant improvement (s.a., recency rehearsal, Ratcliff, 1990; Robins, 1995) or required the entire domain of previously learned items (s.a., random and sweep rehearsal, Robins, 1995). Note that there are alternative approaches to mitigating CI such as modifying the hidden unit representations (node sharpening, French, 1991, and context biasing, French, 1994, and sparse hidden unit representations, Kruschke, 1992) however, following McClelland, McNaughton and O'Reilly, in light of observed psychological phenomena (s.a., temporally graded retrograde amnesia, see McClelland, McNaughton & O'Reilly, 1994, for a review), we chose to investigate interleaving as the primary method for mitigating CI.

Much of the past work on CI has been performed using domains that lack inherent structure. By contrast most cognitive domains are highly structured (e.g., words are composed of letters, speech of phonemes). Domains comprising events that are represented as multiple components are termed *combinatorial domains*. The regular structure of combinatorial domains is fundamental to the productivity of cognitive systems (e.g., the ability to produce an unlimited number of novel words from a finite set of letters), but, somewhat surprisingly perhaps, purely regular domains are rare, and most cognitive domains contain exceptions to the regular structure (forming "quasi-regular" domains, such as the pronunciation of English words). MLPs trained on combinatorial data have been shown to possess very different properties than those trained on random data (Brousse & Smolensky, 1989; Phillips & Wiles, 1993), in particular demonstrating high levels of generalisation, and some of these findings have been related to a reduction in CI (Brousse & Smolensky, 1989). Due to the cognitive relevance of combinatorial domains and the success of MLPs in learning combinatorial structure, we have studied both regular and quasi-regular combinatorial domains.

Results to date

Previous simulations of the A-B and recency rehearsal tasks[1] (Gray & Wiles, 1996) demonstrated that an MLP can extract the regular structure behind regular and quasi-regular combinatorial domains and that the structure of the data itself mitigates interference to a level that could not be called catastrophic. This result is consistent with expectations from other studies (Brousse & Smolensky, 1989), however, in our simulations not even large batch sizes eliminated all interference for quasi-regular domains. After irregular items left the current training batch, performance on these items was retained only briefly and then lost permanently. As irregular items were added to the training batch, performance was lost on similar regular items (two letters in common) not present in the training batch.

[1] In the A-B task, the MLP is trained on a data set (Set A) until all items are learned, followed by training a second set (Set B) without continuing presentations of Set A. This form of A-B Task is common in catastrophic interference literature - items within each data set are trained using an interleaving process, but no interleaving occurs between the two sets. The Recency Rehearsal Task differs from the A-B Task in that after training the initial Set A to criterion, new items are added and older items removed from the training batch incrementally until all items in the domain have been trained.

Aim of Present Work

The current study delved further into the interference effects on MLPs constrained by a limited training batch size. Since our earlier studies showed that CI was not an issue for quasi-regular data, our focus was on the interference that did occur, which we explored by separating the performance of regular and irregular items.

A new rehearsal scheme was designed, similar to recency rehearsal in that after training set A to criterion, new items are added and older items removed from the training batch incrementally. The schemes differ in that in recency rehearsal, the oldest item was removed, whereas in error rehearsal, each item was evaluated with respect to the MLPs performance on the item, and the best learned item (over several trials) was removed.

We hypothesised two ways by which the scheme could improve performance: The first derives directly from the maintenance of less-well-learned items for longer in the training batch, allowing increases in training time and chance of interleaved training with new similar items. The second aspect is more subtle, but directly addresses the critical question of avoiding interference on items that are no longer present in the current training batch (which we call "shielding"): Irregular items entering the training batch would have high error and interfere with the systems performance on similar regular items, increasing their error also. These irregular and similar regular items would then be concentrated in the training batch as the highest error items. Subsequent interleaved presentations would facilitate their separation in HU space, possibly to the extent of shielding the irregulars (after they have left the training batch) from interference by later items.

Method

Simulations were run to compare the new error rehearsal scheme with recency rehearsal. Three measures were of interest: firstly, how well the structure of regular and irregular items is incorporated overall into the HU space (a domain performance measure); secondly, performance on regular and irregular items from the initial batch at the end of training all items in the domain; and thirdly, the degree to which items that are not in the final training batch are maintained (a shielding measure).

For both recency and error rehearsal schemes, changes to the simulations from the previous work (Gray & Wiles, 1996) involved doubling the number of HUs to ensure that capacity was not limiting the performance of the rehearsal schemes, and increasing the maximum number of epochs between training batch updates. All other factors (i.e, data sets and parameters) remained the same (these are reproduced below for completeness).

Structure of the quasi-regular domain

The data used was from an artificial data set designed by Sally Andrews at the University of New South Wales (personal communication) to reflect effects in mapping three letter syllables to their phonetic pronunciation. The input is formed by combining three letters (the onset, vowel and coda). Each of the three letter positions may adopt one of six letters. For example, the onset is one of the letters B, C, D, G, H or S. The mapping was quasi-regular, in which the outputs for the onsets and codas were identically mapped, but vowels were mapped to one of two possible phonetic representations depending on whether they were long or short. Two of the six vowels were always mapped to their short phonetic representation, with the other four varying depending on the combination of the onset and coda letters. Of the 216 (6x6x6) syllables in the quasi-regular domain, 204 adopted the short phonetic representation, and 12 adopted the long (i.e., giving 5.6% exceptions). The exceptions in the domain were randomly distributed with no underlying structure determining which inputs formed the exceptions.

Structure of the MLPs

To represent the domains for training an MLP, each of the three letter positions of the input was represented by a six-bit local code with one unit active per letter. Combined, the input vector thus contained 18 units, with three units active per syllable. Similarly for the output, each letter's phonetic representation was represented by a local code, resulting in 6 units each for the onset and coda, and 10 units for the vowel. Combined, the output vector thus contained 22 units with three units active for any syllable. 36 hidden units were used, forming an 18-36-22 feedforward MLP.

Training batch sizes

The underlying structure of a domain is revealed through the items in the training batch, and hence the size of the batch is critical to the amount of information available for the network to learn. We tested three batch sizes: a small size of eight (the batch size used by McCloskey & Cohen, 1989), a medium size of 50 (used by Brousse & Smolensky, 1989) and a large batch size of 108 (50% of the entire data set).

Training procedure

The considerations for three batch sizes in both recency and error rehearsal schemes resulted in 6 conditions to test (3 batch sizes x 2 schemes). For each condition, ten replications were run, each with a different training batch and random initial weights. Training was via backpropagation with weights updated after every pattern presentation (parameters were as used by McCloskey and Cohen (1989), i.e., learning rate 0.1, momentum 0.9, targets 0.1 and 0.9 and initial weight range [-0.3, 0.3]) and continued until all output units were within 0.2 of their corresponding targets for all patterns in the batch or the maximum number of epochs between training batch updates exceeded (18, 46 and 76 epochs for the small medium and large training batch sizes respectively). The initial training batch was randomly selected, as were the items incrementally added (using a non-replacing selection scheme).

Results and Discussion

Results for recency rehearsal

The recency rehearsal results were qualitatively similar to our previous study (see Figures 1-3), showing that recency rehearsal allows all of the batch sizes (even the smallest) to learn the underlying regular structure of the domain. This finding replicates the earlier one that interference on regular structure is not a major concern for quasi-regular domains.

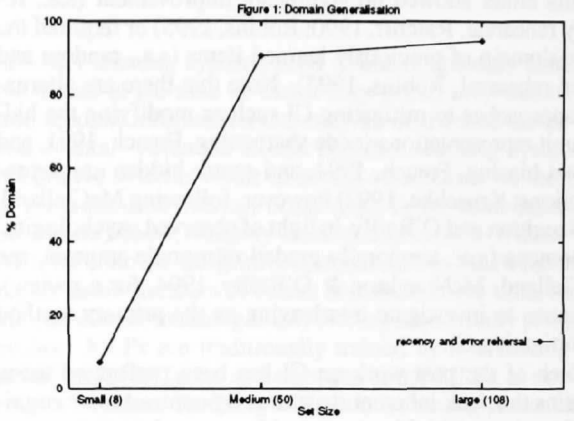

Figure 1: Generalisation to the domain after training the initial batch. This performance characterizes both recency and error rehearsal simulations as the behaviour of the rehearsal schemes is no different at this stage. The small batch size enabled little generalization, indicating that the network had not learned the underlying regular structure behind the domain. Both the medium and large batch sizes enabled the majority of the structure to be learned.

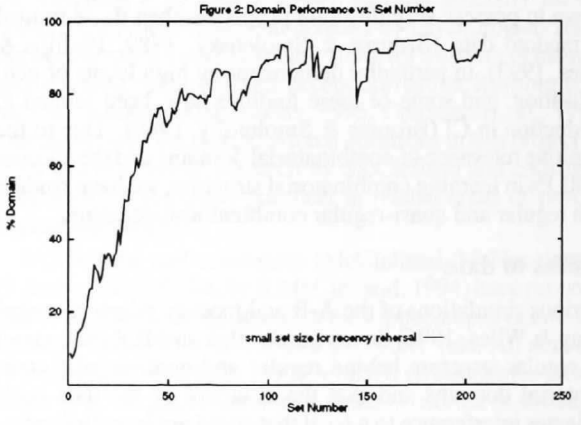

Figure 2: Performance of the domain vs set number for one replication of the small batch size. The graph is qualitatively representative of the behaviour of all replications, showing that immediately after training Set A, the network has learned little of the underlying structure of the domain. As additional items are added to the training batch, performance increases steadily, demonstrating the ability of the MLP to extract structure even from the small batch sizes. This phenomena replicates that observed in earlier simulations using 18 HUs.

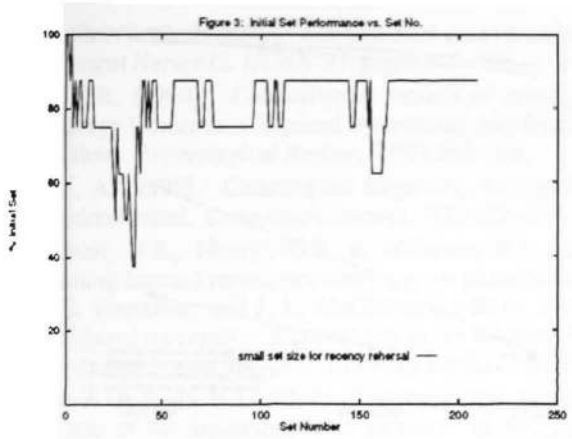

Figure 3: Performance of Set A vs Set Number for the small batch size (same replication as Figure 2). Immediately after training Set A, all items are perfectly learned. Further training results in interference as observed by the drop in performance. The striking aspect of this graph is that as training continues and the MLP learns the underlying structure of the domain, performance on these items returns to a high level. The classic U-shape was observed in all replications.

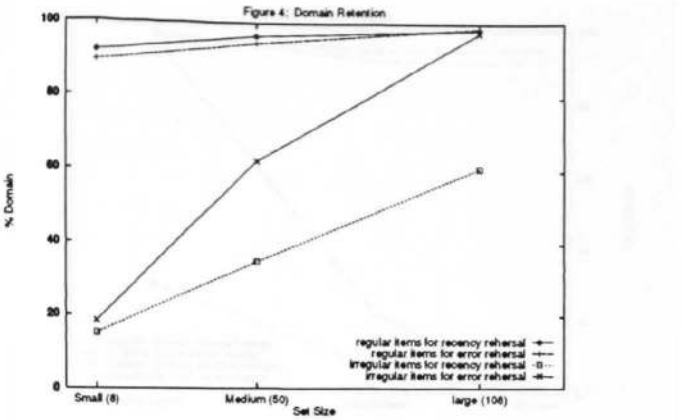

Figure 4: Final performance on the entire domain for both recency and error rehearsal. For both rehearsal schemes, best performance was obtained for the large training batch size (recency - 96.2%; error rehearsal 98.7%). Both schemes show an ability to incorporate irregular items into the regular structure to some degree, however, error rehearsal has incorporated a much higher percentage of the irregular items.

For this simulation, we explored the basis of the performance for the large training batch, as it displayed the best performance: With 36 HUs, recency rehearsal is able to capture the regular structure (98.3% regular items correct), and incorporate some of the irregular structure (60.8% irregular items correct, see Figure 4). We further divided the domain into items present in the training batch, and those that had been dropped. The structure developed in the HUs extended to regular items no longer in the training batch (96.6% correct, see Figure 6) indicating substantial shielding. However, for irregular items that were no longer in the training batch, performance decreased markedly (28.7% correct), showing that recency rehearsal did not shield irregular items from interference after they left the training batch.

With respect to McClelland, McNaughton and O'Reilly's (1994) theory of the H/NC, this result demonstrates that recency rehearsal would not be a suitable model for the interleaving process as it implies that memories that conflict with the majority of the structure in the neocortex would have a low probability of being maintained in the neocortex longer than their duration in the hippocampus.

Comparison of recency and error rehearsal schemes

On all three measures of performance (domain retention, set one retention and shielding - see Figures 4, 5, and 6 respectively) the error rehearsal results were similar to recency rehearsal for regular items, but were markedly improved for irregular items for the medium and large batch sizes. The results indicate two contributing factors - retention of items and shielding:

For the error rehearsal scheme, the domain and set one retention for the large batch size (see Figures 4 and 5) show that regular structure has been maintained with almost all irregular items integrated correctly. These results are understandable given the way in which the error rehearsal scheme retains difficult-to-learn items in the training batch: The error rehearsal scheme concentrated a far higher proportion of irregular items than the initial random distribution, especially for the large training batch (90.0 % of irregular items were retained in the final training batch). Subsequent performance on irregular items is not surprising from a connectionist point of view, but nonetheless may have relevance for cognitive systems as it indicates a method to separate items into functionally regular and irregular, without prior need to define which items are which. It provides one mechanism by which a limited capacity memorisation system could deal with the problems of continually being presented with new items, although retention of items alone is not consistent with the view of the hippocampus as an intermediate term store.

The second contributing factor is due to shielding of items: Figure 6 shows the performance of items that were correct at the end of the simulation but no longer in the training batch (they are the traditional measure in the CI literature). In error rehearsal, a majority of the irregular items were shielded from later interference in the large batch size (58.2% correct compared to 28.7% for recency rehearsal, see Figure 6).

Together the retention of irregular items and shielding provide support for error rehearsal as a suitable model for the interleaving process. Whether such performance is sufficient, or could be further improved is now a feasible question.

There is no direct comparison that can be made with the human data for this type of rehearsal task. Barnes and Underwood (cited by McCloskey & Cohen, 1989) reported retroactive interference of 52% on the AB-AC task, and McCloskey and Cohen (1989) defined CI as substantially lower than this value. At a criterion of 52%, the error rehearsal scheme could be said to be shielding the irregular items no longer in the training batch from CI. Further studies would be required to test the robustness of this result for a variety of quasi-regular domains to have confidence in the numeric value per se, but the results demonstrate a proof of concept.

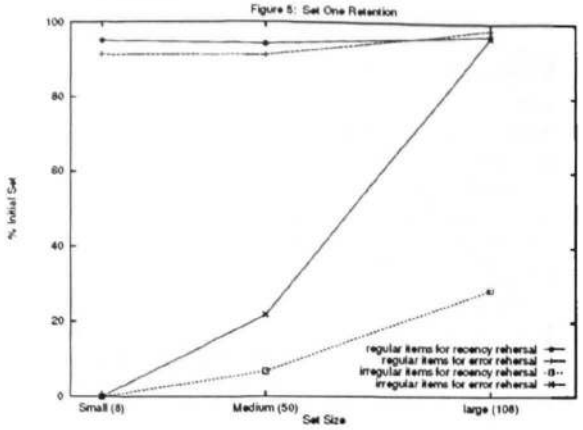

Figure 5: Final retention of the initial batch for both recency and error rehearsal. For both rehearsal schemes, the best performance is achieved for the large batch size (error rehearsal - 98.1%; recency - 92.4%). The results indicate that most of the regular items from set 1 have been retained for all three batch sizes. As in the domain performance (Figure 4), there is a marked difference between the two training schemes for the performance on the irregular items: In particular, for the large batch size on irregular items, error rehearsal gave excellent retention (96.2%) whereas recency rehearsal performance could be considered catastrophic (28.7%).

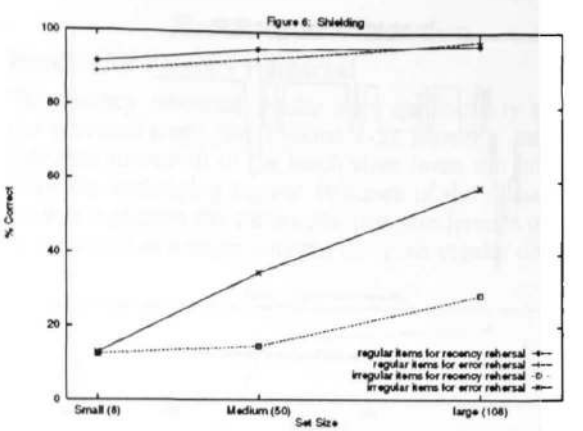

Figure 6: Shielding of items not present in the final training batch. Regular items are maintained by both schemes. For irregular ones, both schemes are able to maintain some items for the large batch size, but recency rehearsal would be considered catastrophic (28.7%) whereas error rehearsal maintains more than half the irregularities (58.2%).

In conclusion, the simulations have shown that the error rehearsal scheme displays promising ability to not only integrate irregular items into the regular structure of the MLP with a limited training batch size, but also shield items from interference after leaving the current training batch, consistent with theories of the hippocampus as both a finite capacity and an intermediate term memory store.

Acknowledgements

We thank Sally Andrews for use of the data set. This work was partially supported by an ARC grant to the second author.

References

Bakker, P.E (1995). *On the implementation of quasiregular mappings by feedforward connectionist networks.* PhD thesis, Departments of Computer Science and Psychology, The University of Queensland.

Brousse, O. & Smolensky, P. (1989). Virtual memories and massive generalization in connectionist combinatorial learning. In *Proceedings of the 11th Annual Conference of the Cognitive Science Society*, pages 380–387.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Model of reading aloud: Dual route and parallel-distributed-processing approaches. *Psychological Review*, 100(4):589–608.

Dennis, S.J. (1994). *The integration of learning into models of human memory.* PhD thesis, Department of Computer Science, The University of Queensland.

French, R.M. (1991). Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks. Technical Report 51-1991, Center for Research on Concepts and Cognition, Indiana University, 510 North Fess, Bloomington, Indiana 47408.

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.*

Gray, B. & Wiles, J. (1996). Interference in regular and quasi-regular combinatorial domains: Insights from a two-component memory system. In *Proceedings of the Seventeenth Australian Conference on Neural Networks.* To appear in.

Humphreys, M.S., Bain, J.D. & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, 96(2):208–233.

Kruschke, J.K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review*, 99(1):22–44.

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. Technical Report PDP.CNS.94.1, Dept. of Psychology, Carnegie Mellon University, ftp://hydra.psy.cmu.edu:/pub/pdp.cns.

McCloskey, M. & Cohen, N.J. (1989) Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation*, 24:109–165.

O'Reilly, R.C. & McClelland, J.L. (1994). Hippocampal conjunctive encoding, storage and recall: Avoiding a tradeoff. Technical Report PDP.CNS.94.4, Carnegie Mellon University.

Phillips, S. & Wiles, J. (1993). Exponential generalizations from a polynomial number of examples in a combinatorial

domain. In *Proceedings, International Joint Conference on Neural Networks, IJCNN'93*, pages 505–508.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, 97(2):285–308.

Robins, A. (1995). Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel distributed processing: Explorations in the microstructure of cognition*, pages 318–362. Bradford Books / MIT Press.

Treves, A. & Rolls, E.T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus*, 4(3):374–391.