This is a pre-publication draft. Please cite the published version:
Robins, A. Catastrophic forgetting, rehearsal, and pseudorehearsal. Connection Science: Journal of Neural Computing, Artificial Intelligence and Cognitive Research, 7 : 123 - 146 (1995)

Catastrophic Forgetting, Rehearsal, and Pseudorehearsal

Anthony Robins
Computer Science Department
University of Otago, P.O Box 56, Dunedin
New Zealand

email: coscavr@otago.ac.nz
Ph: +64 3 4798578

Running heading: Catastrophic forgetting

Keywords: Catastrophic forgetting, catastrophic interference, stability, plasticity, rehearsal

Abstract

This paper reviews the problem of catastrophic forgetting (the loss or disruption of previously learned information when new information is learned) in neural networks, and explores rehearsal mechanisms (the retraining of some of the previously learned information as the new information is added) as a potential solution. We replicate some of the experiments described by Ratcliff (1990), including those relating to a simple "recency" based rehearsal regime. We then develop further rehearsal regimes which are more effective than recency rehearsal. In particular "sweep rehearsal" is very successful at minimising catastrophic forgetting. One possible limitation of rehearsal in general, however, is that previously learned information may not be available for retraining. We describe a solution to this problem, "pseudorehearsal", a method which provides the advantages of rehearsal *without* actually requiring any access to the previously learned information (the original training population) itself. We then suggest an interpretation of these rehearsal mechanisms in the context of a function approximation based account of neural network learning. Both rehearsal and pseudorehearsal may have practical applications, allowing new information to be integrated into an existing network with minimum disruption of old information.

1 Introduction

Despite their many successes in recent years, neural networks are not without their limitations. Some of these are quite theoretical, such as problems and open questions related to the compositionality of distributed representations; and some are practical, such as the inability of neural networks to "explain" the method or reasoning by which a given output is produced.

One practical problem which is very pervasive is the "stability / plasticity dilemma" (see for example Grossberg (1987), Carpenter & Grossberg (1988)). Ideally the representations developed by neural networks should be *plastic* enough to change to adapt to changing environments and learn new things, but *stable* enough so that important information is preserved over time. The dilemma is that while both are desirable properties, the requirements of stability and plasticity are in conflict. Stability depends on preserving the structure of representations, plasticity depends on altering it. An appropriate balance is difficult to achieve.

One consequence of a failure to address stability / plasticity issues in many neural networks is excessive plasticity, often somewhat dramatically labelled "catastrophic forgetting" (or "catastrophic interference"). The problem of catastrophic forgetting can be summarised as follows:

> If after its original training is finished a network is exposed to the learning of new information, then the originally learned information will typically be greatly disrupted or lost.

Carpenter & Grossberg (1988) suggest the analogy of a person growing up in one city, and then later moving to a second city. Catastrophic forgetting would be if the process of learning about the second city caused the person to forget what they knew of the city they grew up in. Catastrophic forgetting is implausible an aspect of cognitive models, human memory does not typically suffer from this problem! It is also very undesirable in practical terms, making it very difficult to modify or extend any given neural network application.

A number of recent studies have highlighted the problem of catastrophic forgetting and explored various issues, typically in the family of back-propagation type networks – these include McCloskey & Cohen (1989), Hetherington & Seidenberg (1989), Ratcliff (1990), Lewandowsky (1991), French (1992, 1994), McRae & Hetherington (1993), Lewandowsky & Li (1994), and Sharkey & Sharkey (1994a, 1994b). Some authors have also noted the problem in the context of Hopfield networks – these include Nadal, Toulouse, Changeux & Dehaene (1986), and Burgess, Shapiro & Moore (1991).

While stability / plasticity issues are very general, the term "catastrophic forgetting" has tended to be associated with a specific class of networks, namely static networks employing supervised learning. This broad class includes probably the majority of commonly used and applied networks (such as the very influential back-propagation family and Hopfield nets as noted above). Other types of network – for example dynamic / "constructive" networks and unsupervised networks – are not necessarily prone to catastrophic forgetting as it is typically described in this context. Dynamic networks are those that use learning mechanisms where the number of units and connections in the network grows or shrinks in response to the requirements of learning (see examples reviewed in Hertz, Krough, & Palmer (1991)) – new units can be added to encode newly learned information without disrupting existing units. In unsupervised networks there are no "correct" / target outputs to be forgotten. More general stability / plasticity issues certainly arise, but may be more amenable to solution in the unsupervised framework – Gillies (1991) provides an excellent discussion of these issues and explores possible methods.

The starting point for this paper is some of the experiments reported by Ratcliff (1990) on catastrophic forgetting in back-propagation networks. In Section 2 we introduce and replicate the relevant experiments regarding catastrophic forgetting and the use of a rehearsal mechanism as a potential solution. (Rehearsal involves retraining previously learned information as new information is introduced). In Section 3 we explore further rehearsal mechanisms, and describe "sweep rehearsal", a very effective method. One limitation on the use of rehearsal, however, is that previously learned information may not be available for retraining. Section 4 addresses this problem. A new method, "pseudorehearsal", is described which is relatively effective at providing the benefits of rehearsal without requiring access to the original information on which the network was trained. In Section 5 we propose an interpretation of rehearsal mechanisms within the context of a function approximation based account of neural network learning – predictions about the effectiveness of pseudorehearsal are suggested by this interpretation. Practical issues and the application of the new rehearsal mechanisms to real world populations are discussed in Section 6.

2 Catastrophic forgetting and rehearsal in a back-propagation network

2.1 Background

Ratcliff (1990) presents an extended demonstration and exploration of catastrophic forgetting in a back-propagation network. We have duplicated Ratcliff's results using exactly the same architecture and training methods that he did (see Robins (1993)). For consistency with later simulations in this paper, however, in this section we chose to illustrate the same points using slightly different methods, as described below.

Experiments are based on the use of a 32-16-32 standard back-propagation network (Rumelhart, Hinton & Williams (1986)) using on-line training (weight update after every pattern). The learning rate is set to 0.3 and the momentum constant to 0.5. A population is considered to be learned when the output unit activation for every unit is within 6% of its target value (for each item in the population). These are the same parameter settings and error criterion as used by Ratcliff. This network is used to learn a *base population* of items (input/output pairs of binary vectors where vectors are constructed by setting elements to 0 or 1 at equal probability) in the usual way. Subsequently a number of further items are learned one by one, these are the *intervening trials.* The *goodness* of the base population can be plotted at any time to show the effect of the intervening trials on the networks ability to correctly recall the base population output vectors[1].

Within this framework the methods of the experiments reported in this section vary slightly from those of Ratcliff, although equivalent results are produced. Ratcliff used an "encoder" or "autoassociation" training paradigm. We have duplicated all experiments using both *autoassociation* (the target output vector is identical to the input vector) and *heteroassociation* (the target output vector is independent of the input vector), where all vectors are constructed at random as described above. For brevity we report only heteroassociative results until the summary (Section 5) where autoassociative results are also shown. Ratcliff used a mixture of network sizes, employed differing numbers of items in various experiments, and terminated training in various experiments using either an error criterion or a fixed number of epochs. Throughout this paper we report only experiments using the larger of Ratcliff's networks (32-16-32), we use base populations consisting of 20 items (I/O vector pairs) with 10 intervening trial items, and we always terminate training using the specified error criterion.

Finally, in presenting results Ratcliff uses a mixture of graphs including *base population goodness* graphs showing the (mean) goodness of the base population after varying numbers of intervening trials; and *serial position* graphs showing the goodness of each learned item (base and intervening) individually after all training has been completed. Again for brevity we here illustrate results using only base population goodness graphs until the summary (Section 5) where serial position results are also shown. In all figures in this paper the graphs shown plot the results of averaging the data from 50 replications of the given experiment, identical except that a different randomly constructed population is used for each replication. In other words all graphs show average results for 50 different populations.

2.2 Catastrophic forgetting

[1] "Goodness" is a measure of the networks ability to correctly reproduce the appropriate outputs for a given set of inputs. It is calculated by averaging for each input / output pair the normalised dot product of the target vector and the actual output vector observed (see Ratcliff, 1990, p288). Vectors are transformed so that a goodness value of 1 indicates a perfect match and a value of 0 indicates chance performance (50% match).

Catastrophic forgetting is a complex phenomena involving many practical considerations. In this section we replicate Ratcliff's (1990) simple demonstration of the effect, deferring a discussion of the more complex issues until Section 6. The basic demonstration involves inspecting the (mean) goodness of a base population after some number of intervening trials. Ratcliff used base populations of 1 item and 4 items, we include a base population of 20 items (for consistency with most other simulations in this paper). In each case the base population is trained using back-propagation in the usual way as described above (Section 2.1) A further 10 intervening trial items are then learned one after another, and the average goodness of the base population patterns calculated after each intervening item.

Figure 1 shows the fall the base population goodness for the three base population size conditions. (Recall that all figures show results for each condition averaged over 50 replications using different populations). After even a single intervening trial the ability to correctly reproduce the base population outputs is significantly disrupted. Ratcliff variously describes this drop in goodness, which he replicated over a number of conditions, as "serious" or "massive". He also noted (Ratcliff 1990, p291) that items learned in groups (i.e. the 4 item base population and in our simulation the 20 item base population) were significantly more resistant to forgetting.

2.3 Recency Rehearsal

Ratcliff went on to experiment (among other manipulations which are beyond the scope of this paper) with the use of a simple rehearsal mechanism for reducing the catastrophic forgetting effect. Rehearsal is the relearning of a subset of previously learned items at the same time that each new item is introduced. It is reasonable to suppose that this process could "strengthen" the previously learned items or "protect" them from disruption by new items, and take advantage of the robustness of training items in groups (noted above). Accordingly, Ratcliff modified the training paradigm so that new intervening items were introduced one at a time and trained in a buffer along with the three most recently learned items. In other words the buffer consists of a queue of four items where the oldest item is dropped out as each new item is introduced. Each buffer is trained fully (until all items in the buffer are trained to criterion) before the next item is introduced and a new buffer created.

In our simulations the base population (1, 4 or 20 items) was trained using back-propagation in the usual way as described above (Section 2.1). Additional items were then added one at a time and trained in a buffer of the four most recent items (as described above). The results, as illustrated in Figure 2, confirm that this form of rehearsal has a very modest impact on the catastrophic forgetting effect. The base population is still significantly disrupted, falling to a goodness of around 0.3 after 10 intervening items, although the drop in goodness is not as sudden as in the case of no rehearsal (cf Figure 1)[2]. The rehearsal mechanism is, however, significantly better than no rehearsal at maintaining the goodness of the intervening trial items, especially those learned towards the end of the sequence (see serial positions 11 to 20 in the "recency" condition of the serial position graphs Figures 8 and 10 — these are described in the summary, Section 5). Ratcliff (1990, p294 – 295) analyses the behaviour of this form of rehearsal under different conditions.

To summarise, if a back-propagation network has learned a population of items (the base population), then the learning of any subsequent items (intervening trials) significantly disrupts the ability of the network to correctly reproduce the output vectors of the base population. Rehearsing the three most recently learned items as each new item is added to the network does not significantly improve this situation.

3 Rehearsal

3.1 Introduction

Ratcliff's recency rehearsal mechanism retrains the three most recently introduced items as each new intervening item is introduced. Very few base population items are

[2] Note that as the first three intervening patterns are added it is necessary to select some members of the training buffer from the base population, but that from the fourth intervening trial on the buffer consists solely of the most recent intervening patterns. The initial increase in the goodness of the one item base population is an artifact of this mechanism – for the first three intervening trials the whole of the base population (a single pattern) remains in the training buffer.

ever included in the rehearsal process (none at all after the first three intervening trials) so the goodness of the base population is not maintained.

Using the same architecture and overall training paradigm we have experimented with a number of different rehearsal regimes. A regime is defined as a method of choosing the three previously learned items (base population or intervening) to be included in the rehearsal buffer with each newly introduced intervening trial item. As in recency rehearsal the items in the rehearsal buffer are trained (or "retrained" in the case of the three previously learned items) fully to criterion. We have explored a range of alternative regimes. By using strategies which can include any previously learned item, including base population items, it was hoped that the goodness of the base population could be maintained.

Ratcliff's recency rehearsal is based on a "simplified model" of human rehearsal in list learning (Ratcliff, 1990, p293). We do not claim that the alternative regimes that we explore can be directly related to human performance, the goal is simply to find the most effective regime.

3.2 Random rehearsal

We have explored a range of new regimes that can include any previously learned item in rehearsal: choosing three items at random, the three items most similar to the new intervening item, and the three items least similar to the new intervening item. In practice there was no difference in the performance of these variations — the results were indistinguishable from each other. Consequently for the purposes of this paper we describe only the simplest option, the choice of three items at random.

Figure 3 shows the effect of this "random" rehearsal for a base population of 20 items. While the base population is still disrupted, it is maintained at a significantly better goodness than is achieved by recency rehearsal. In terms of serial position, random rehearsal is better at maintaining the the goodness of base population items than intervening items (see the "random" condition of the serial position graphs, Figures 8 and 10 — these are described in the summary, Section 5). A fall in goodness of intervening items introduced early in the sequence is characteristic of many of the rehearsal mechanisms described in this paper, see Section 6.2 for further discussion.

3.3 Sweep rehearsal

Although the standard regimes which can include any previously learned items as typified by random rehearsal generally improved performance as described above, one different kind of regime clearly stood out as yielding exceptionally good performance for the whole population. We called this regime "sweep" rehearsal.

Sweep rehearsal is based on the use of a "dynamic" training buffer rather than the fixed training buffer used in the standard regimes. Intervening items are introduced one at a time and trained in a buffer with three previously learned items as usual, however the three previously learned items are chosen at random *for each epoch.* Training progresses over a number of epochs until the new item, which is always in the buffer, is trained to criterion. Note that compared to standard rehearsal mechanisms such as random rehearsal, sweep rehearsal will expose *more* previously learned items to one or more training epochs, but does *not* actually specifically retrain any previously learned item to criterion.

This regime yields excellent results. As shown in Figure 4 the base population is maintained at a very high level of accuracy – in fact ability to reproduce the correct output vectors actually *improves* slightly over the intervening trials. (Intervening items are also maintained at a high level of accuracy – see Figures 8 and 10 in the summary, Section 5).

Sweep rehearsal remains effective as the number of intervening trials is significantly increased, although eventually of course limitations are encountered. In the case of heteroassociative learning the goodness of learned items deteriorates gradually, but steadily as more and more intervening items are added. In the case of autoassociative learning the goodness of the learned items deteriorates even more gradually. We found little drop in performance in trials using up to 70 intervening items. This result is deceptive, however, as the high goodness of the learned items masks an underlying problem arising from the autoassociative learning task. This problem is the networks poor ability to distinguish between the learned items and novel items (a newly generated test population). In other words, while the goodness of the learned items remains high, the extent to which they are genuinely learned (able to be distinguished from novel items) is significantly less than in the heteroassociative case. These results are presented in more detail in Appendix A.

3.4 Discussion

The success of random and sweep rehearsal mechanisms is hardly a surprising result. Both methods expose previously learned items to further training, maintaining the goodness of these items as new intervening items are introduced. What is surprising, perhaps, is the marked superiority of the sweep rehearsal approach.

Random rehearsal retrains 3 previously learned items to criterion for every new intervening item introduced. Sweep rehearsal is more successful at maintaining a high goodness for previously learned items despite the fact that it never retrains them to criterion. Instead, for every new item introduced sweep rehearsal exposes several previously learned items to just one or more presentations / weight updates. In general terms, then, the “broad and shallow” approach of the sweep regime is a much more successful rehearsal strategy than the “narrow and deep” approach of the random regime.

We suggest, however, that the limits that the sweep rehearsal approach encounters as the number of intervening trial items is increased are manifestations of fundamental upper limits on the performance of rehearsal mechanisms in general that arise from the constraints imposed by finite network architectures. While an effective rehearsal mechanism will allow large numbers of additional items to be successfully added, these constraints will eventually manifest themselves in falling goodness and / or discriminability (see Appendix A for further discussion).

4 Pseudorehearsal

4.1 Introduction

While sweep rehearsal is a very successful approach, it could be argued that the access that it assumes to all previously learned items makes the method of limited interest. Indeed our own experiments have shown that to integrate new items with old / previously learned, there is only slight advantage (in terms of speedup of training) in using sweep rehearsal over simply retraining the network in the usual way from scratch on an extended population containing all the desired items.

In this section we shall show, however, that it is possible to add intervening items to a previously trained base population (as described in the previous sections) using rehearsal mechanisms *even without the use of any previously learned items.* Furthermore, sweep rehearsal still performs well. This approach, which we have called “pseudorehearsal”, therefore provides a method for integrating new information into a network without requiring any access to the population on which the network was originally trained.

4.2 The pseudorehearsal mechanism

Pseudorehearsal is based on the use in the rehearsal process of artificially constructed populations of "pseudoitems" instead of the "actual" previously learned items. A pseudoitem is constructed by generating a new input vector (setting at random 50% of input elements to 0 and 50% to 1 as usual), and passing it forward through the network in the standard way. Whatever output vector this input generates becomes the associated target output. (Note that using standard back- propagation these output vectors will contain real values instead of the binary values used in the actual items).

A population of pseudoitems constructed in this way can be used instead of the actual items in rehearsal for any of the standard regimes, or the sweep rehearsal regime, described above. The given regime proceeds as usual, except that whenever it is necessary to chose a previously learned item or items for rehearsal, a pseduopoulation is constructed and a pseudoitem or items are chosen instead. More detailed descriptions of the specific examples of random and sweep pseudorehearsal are presented below.

As we shall show below, this pseudorehearsal is reasonably effective. The pseudoitems serve as a kind of "map" of the function appropriate to reproducing the actual population, and by using them in the rehearsal process as a framework for integrating the new intervening trial items the function appropriate to the actual population is preserved (see Section 5.2 below). If this interpretation is correct then we would expect that the larger the population of pseudoitems used in rehearsal the better the map of the appropriate function, and the more effective the pseudorehearsal at preserving the goodness of the actual base population items. The following simulations were designed to test this prediction and explore the effectiveness of random and sweep pseudorehearsal.

4.3 *Random pseudorehearsal and sweep pseudorehearsal*

In the simulation of random pseudorehearsal we start with a network trained on a base population of 20 items, and wish to introduce 10 intervening trial items in the usual way. Assuming that the base population is not available for rehearsal, we generate a pseudopopulation as described above. The first intervening trial item is introduced and trained in a training buffer with three randomly selected pseudoitems (as for random rehearsal described in Section 3.2) until the error criterion is reached. This process (including the generation of a new pseudopopulation) is repeated for the each intervening item, until all intervening items have been learned.

The simulation of sweep pseudorehearsal proceeds in the same way except that the dynamic training buffer of sweep rehearsal (as described in Section 3.3) is used. Each intervening trial item is trained to criterion in a buffer where the three other items are chosen at random each epoch from the pseudopopulation.

Figures 5 and 6 show the results of these simulations for random and sweep pseudorehearsal respectively, using pseudopopulations of size 8, 32, and 128. The goodness of the original / actual base population is plotted after each intervening trial as in previous figures. Random pseudorehearsal (Figure 5) is only moderately effective at preserving the goodness of the base population. The size of the pseudopopulation no effect on the efficacy of rehearsal – this is to be expected as although selecting from a larger pool of possible rehearsal pseudoitems, within any given buffer only three pseudoitems are used. As expected from earlier rehearsal results, sweep pseudorehearsal (Figure 6) is significantly more effective than random pseudorehearsal. Furthermore the size of the pseudopopulation has a clear impact on its efficacy. This is consistent with our predictions, noted above, as the dynamic buffer used in sweep rehearsal uses a large number of pseudoitems within each of the ten rehearsal buffers, fully exploiting a larger pool of possible rehearsal pseudoitems. In short, larger populations of pseudoitems allow more pseudoitems to be included in rehearsal, resulting in a more accurate preservation of the function appropriate to reproducing the actual base population.

5 Rehearsal and pseudorehearsal: Summary and interpretation

5.1 Summary: Performance of the regimes

We have described several forms of rehearsal mechanism as possible solutions (or part solutions) to the catastrophic forgetting problem. The following figures summarise and extend these results.

Figure 7 draws together the results so far presented above. The graph shows the base population goodness after each intervening trial for the various forms of rehearsal mechanism including no rehearsal, recency rehearsal, random rehearsal, sweep rehearsal, random pseudorehearsal, and sweep pseudorehearsal. (Note that only base populations of size 20, and pseudopopulations of size 128 are plotted). Sweep rehearsal is clearly the best performer, actually increasing the goodness of the base population as intervening items are added. The next best performance is by sweep pseudorehearsal (out performing other rehearsal mechanisms even though it uses pseudoitems rather than actual items in rehearsal!). This regime provides the best method of adding further items to a network when the original training data (the actual base population) is not available. Other regimes perform with varying degrees of success as described above.

Figure 8 shows a different way of interpreting the performance of exactly the same regimes. This serial position graph shows the goodness of each individual item (20 base population items and 10 intervening) after all training has been completed. This enables a comparison to be made between the goodness of the base population and the goodness of the intervening items, and highlights any temporal effects of the regimes. Note that all regimes show a drop in goodness for the initial intervening items (item 21 is the first intervening item), the higher base population goodness being consistent with the effects of group training noted in Section 2.2 (see Section 6.2 for further discussion). Most regimes also result in gradually improving goodness for items learned towards the end of training (having progressively less subsequent items to interfere with learning).

The remaining Figures, 9 and 10, show the corresponding base population goodness and serial position graphs for the same regimes using an autoassociative learning paradigm rather than the heteroassociative learning described above. The results are broadly similar. In general the goodness of all items is slightly raised in the autoassociative case. The ordering of the performance of the regimes is the same, with sweep rehearsal still the most effective.

5.2 Interpretation: Function approximation

While we have not carried out any formal analysis of these rehearsal regimes, we suggest that an initial interpretation in the context of function approximation yields some useful insights.

Common "multilayer perceptron" network architectures such as backpropagation can be conceptualised as function approximators (see White (1992) for an overview). Abstracting the behaviour of a network to the "toy" two dimensional example shown in Figure 11(a), the x axis represents the possible inputs to the network i, and the y axis represents the outputs of the network, some function $F(i)$ of the inputs. For a given training population of actual inputs to the network and the actual outputs that they generate, the process of learning is a process of the network fitting some function to these training population data points. A wide range of possible functions will fit the data points – see Sharkey & Sharkey (1994a) for a good discussion of these issues. Depending on the architecture of the network and the details of the training method the actual function learned may be a "compact" function that interpolates well between data points or a "noisy" function which does not (see Figure 11(a)).

It is often desirable for real world tasks to interpolate well (generalise correctly) or to capture an assumed underlying target or "total" function (Sharkey & Sharkey, 1994a) from the training examples. Consequently there are a wide range of techniques which can be applied to constrain a network to learn compact functions (see for example Moody (1994))[3].

[3] To learn a good approximation to an underlying total function (minimise the "prediction risk") it is necessary to

The rehearsal mechanisms described above can be interpreted in this context. Assuming that the network starts with some existing learned function (has learned the base population) then training new intervening items is equivalent to adding further data points and altering the old function so that it fits each new point as it is introduced. With no rehearsal the new learned function is not constrained to continue to fit the old data points and might be significantly different from the old function – see Figure 11(b). This is the cause of catastrophic forgetting, the base population inputs will no longer generate the correct outputs using the new function.

In comparison, rehearsal mechanisms include the old data points in the training process as any new data points are added. This constrains the new function learned by the network to fit both old and new data points (previously learned items including the base population, and new intervening items) – see Figure 11(c). In general terms the new data points are incorporated into the existing function instead of "overwriting" it with a new one – new information is incorporated into the context of the old. Interpreting the success of sweep rehearsal compared with random rehearsal in this context suggests that as each new data point is added it is more useful to constrain the new learned function to fit many old data points approximately rather than a few (three in our simulations) old data points closely.

Pseudorehearsal mechanisms assume that we do not have access to the old data points (base population) to use in this way. Instead the old function is randomly sampled to construct a set of old "pseudodata" points (pseudoitems) to serve the same purpose. Including old pseudodata points in the training process as any new data points are added again constrains the new function learned by the network to fit both old pseudodata and new data points – see Figure 11(d). If the pseudodata points accurately capture the shape of the old function then again the net effect will be to preserve the shape of the old function as much as possible and incorporate the new data points into it. If shape of the old function is well preserved then the base population items will still generate (approximately) correct outputs[4].

---

ensure that the training set samples the underlying function adequately, and that the network fits a compact function to the training set (without overfitting). Moody (1994) provides a useful review of relevant techniques such as cross-validation and sequential network construction.

[4] Note that larger populations of pseudoitems sample the old function more densely and are thus more effective at preserving its shape in the context of sweep pseudorehearsal (see Figure 6).

In this context we can make certain predictions about the efficacy of pseudorehearsal. Pseudorehearsal will not work well in situations where the network fits noisy functions to its data points (consider the example in Figure 11(a)). If after its original training the base population is described by a noisy function, then any constructed pseudodata points will be noisy – not systematically related to the original data points. If training on new intervening items subsequently occurs, then a new (noisy) function will be fit to the (noisy) pseudodata points, compounding the error. Consequently the new function may be a very poor fit to the original data points and the base population items will no longer generate the correct outputs. Conversely, of course, pseudorehearsal will work well in situations where the network is constructed and trained so as to fit compact functions. Information about the base population data points will be systematically preserved in pseudodata points and then in the new function fit to the pseudodata points. In short, pseudorehearsal methods will be most effective in networks which have been constructed to generalise robustly. Note that the method may well be capable of more effective performance than is illustrated by the architecture and randomly constructed populations used in the simulations presented in this paper.

To summarise, we suggest that the rehearsal mechanisms we have described can be usefully interpreted in the framework of a function approximation account of neural network learning. The process of learning the base population items is, conceptually, a process of fitting a function to the data points representing the items. Without rehearsal, adding new data points (training intervening items) results in the learning of a new function which may be unrelated to the old function describing the base population. The various rehearsal mechanisms use either the base population data points, or pseudodata points constructed from the old function, to preserve the shape of the old function as much as possible while accommodating new data points. This interpretation would certainly benefit from a more rigorous and formal analysis.

6 Practical issues

6.1 The complexity of catastrophic forgetting

In our presentation of catastrophic forgetting so far we have deferred discussion of some of the complexities. Catastrophic forgetting does not always occur when new items are added to a network. For example, when new items are added which are further regular instances of a systematic mapping already exhibited by items in the base population then there is a minimal forgetting effect. Techniques for reducing catastrophic forgetting are not relevant in such cases.

Sharkey & Sharkey (1994a, 1994b) provide a useful overview of several practical issues, noting that catastrophic forgetting occurs most significantly in cases where training is sequential and without negative exemplars. They also note that there is an inevitable tradeoff between reducing catastrophic forgetting (increasing generalisation) and the breakdown of "old-new discrimination" (as observed for example in this paper, see the discussion of the discriminability of learned vs novel items in Appendix A).

French (1992) suggests that the extent to which catastrophic forgetting occurs is largely a consequence of the overlap of distributed representations, and can be reduced by reducing this overlap. Several previous explorations of mechanisms for reducing catastrophic forgetting have focused on reducing representational overlap, particularly in the hidden unit representations developed by the network. The novelty rule (Kortge 1990), activation sharpening (French 1992), and techniques developed by Murre (1992), and McRae & Hetherington (1993) all fall within this general framework. French (1994) contains a brief description of most of these methods, and introduces "context-biasing" which produces internal representations that are both well distributed and well separated. (Consequently this method does not suffer from problems that French notes can arise from excessive restriction of the distributedness of representations).

Even if apparently forgotten using the criterion of base population goodness, a "residue" of previously learned items is sometimes still observable in a network in the sense that the previously learned items can be relearned much more quickly than they were learned originally by the network. This speed of relearning is sometimes used as a measure of the efficacy of a proposed mechanism for reducing catastrophic forgetting (see for example French (1992, 1994)).

6.2 Scaling and performance of rehearsal methods

We have presented rehearsal and pseudorehearsal regimes in this paper holding constant, for the most part, the numbers of base population and intervening items, the number of pseudoitems, and the size of rehearsal buffers. Naturally there are a family of questions relating to the behaviour of the regimes under variations in these conditions.

Figure 12 illustrates the behaviour of sweep rehearsal using base population sizes ranging from 10 to 50, with 20 intervening items. Sweep rehearsal continues to perform well over all conditions. Appendix A describes the the use of sweep rehearsal with up to 70 intervening items.

Figure 13 illustrates the behaviour of sweep pseudorehearsal using base population sizes ranging from 10 to 50, with 20 intervening items. Performance of the method falls off with increasing base population size. Our initial experiments have shown that in general a higher goodness can be achieved by increasing the size of the rehearsal buffer. This suggests that the efficacy of this regime is significantly influenced by the size of the buffer in proportion to the size of the base population. An appropriate buffer size can be chosen empirically for a given task according to the desired goodness. In the standard simulations presented in Section 4 the rehearsal buffer includes at best 15% of the previously learned items[5].

There are a number of ways in which the general performance of the regimes that we have described could possibly be improved. Our preliminary experiments have shown that larger rehearsal buffers give better performance, in all regimes the size of the rehearsal buffer could be arbitrarily increased to achieve a desired level of

[5] Three previously learned items in each buffer drawn are drawn from 20 base population items and (as intervening trial items are added) up to 9 previously learned intervening items.

performance. In sweep rehearsal the training of the rehearsal buffer could be extended for an arbitrarily long period (after the new intervening item was trained to criterion) to enable more items to be rehearsed. In sweep pseudorehearsal the number of pseudoitems could be made arbitrarily large, and the training of the rehearsal buffer extended. To improve performance on intervening items, actual intervening items could also be used along with pseudoitems in the rehearsal process, instead of using only pseudoitems at each step as described in this paper. We have also suggested (Section 5.2) that the performance of sweep pseudorehearsal may be affected by properties of the network as a function approximator, and that this regime may well be capable of more effective performance than is illustrated by the architecture and randomly constructed populations presented in this paper.

Finally, recall that in our simulations the goodness of intervening items (especially initial intervening items) is typically somewhat lower than the goodness maintained for the base population (see Section 5.1, Figures 8 and 10). This is characteristic of most of the rehearsal regimes, intervening items are trained singly and do not benefit from the advantages of being trained in a group (see Section 2.2). This situation could be addresses by modifying the training paradigm so that intervening items were introduced in groups, or all at once.

6.3 Real world populations

We have developed the regimes presented in this paper using randomly constructed populations (and replications over 50 different populations in all cases) as this approach seemed to reduce the possibility of the rehearsal method being dependent on any particular structure or regularity in the populations. Systematically exploring the behaviour of these regimes over a range of “real world” populations is the obvious next step, however, which he hope to address in future work.

The function approximation interpretation proposed in this paper provides both reasons to expect that these methods will generalise well, and a framework for interpreting their behaviour. In particular this framework predicts that pseudorehearsal based regimes will be most effective where the network has been constructed and trained so as to generalise well.

7 Discussion

To summarise, there are a number of rehearsal regimes that are more effective than the recency rehearsal described by Ratcliff (1990). Sweep rehearsal is particularly effective, enabling in effect new information to be integrated without disrupting old information at all. Sweep pseudorehearsal provides a mechanism for integrating new information with only moderate disruption of old information, even when the old information (the previously learned patterns on which the network has been trained) is not available for rehearsal. We suggest that function approximation provides a useful framework for understanding and exploring the behaviour of these regimes.

Sweep rehearsal remains quite robust as the size of the base population and the number of intervening items increases. The effectiveness of sweep pseudorehearsal is more dependent on the ratio of the size of the rehearsal buffer to the size of the base population, and also depends on the number of pseudoitems used. These factors can be adjusted to increase the effectiveness of this regime, and we have also proposed a range of other methods which may improve performance. While sweep rehearsal is very effective, there are of course limits on the number of additional items that can be successfully incorporated (see Appendix A). We suggest, however, that the limits encountered by sweep rehearsal are manifestations of fundamental upper limits on the performance of rehearsal mechanisms in general that arise from the constraints imposed by finite network architectures.

Both rehearsal and pseudorehearsal are potential solutions to the catastrophic forgetting problem, allowing new information to be integrated into an existing network. These mechanisms may have practical uses for extending pretrained network applications, however further investigation of their application to real world problems is necessary. It may also be productive to explore the relationship of rehearsal mechanisms to other potential solutions to catastrophic forgetting, in particular investigating the nature of the internal representations developed during rehearsal to see if they tend to minimise overlap (see Section 5.1).

In closing, a fanciful speculation. As noted above, in exploring rehearsal regimes we were not concerned with issues of “psychological plausibility”, but were simply looking for the most effective regime. In retrospect, however, an analogy comes to mind. Sweep pseudorehearsal provides a way of integrating new information with old, using “approximate simulations” of old information (training using pseudoitems). In human beings, one albeit contentious hypothesis about the function of sleep (particularly REM sleep) is that it is a period where newly acquired knowledge is integrated into existing long term memory (see for example Greenberg & Pearlman (1974), Winson (1990)). If our long term memory is accessed via approximate simulations during this process (a function which manifests itself as dreams?) then the analogy with sweep pseudorehearsal becomes clear[6]. Is sweep pseudorehearsal a version of the solution to the catastrophic forgetting problem adopted by the mammalian brain? Is, perchance, a network during sweep pseudorehearsal undergoing the functional equivalent of dreaming?

Acknowledgements

Thanks to the research assistants who assisted with this study, particularly Andrew Gillies (who was instrumental in developing sweep rehearsal) and Nick Mein (who developed the discriminability graphs presented in Appendix A, and for useful discussions). Thanks also to Dr Noel Sharkey for his correspondence regarding sweep rehearsal (independently developed at his laboratory) and catastrophic remembering.

[6] See Winson’s discussion of dreams as the forebrain’s “best fit” interpretation of random PGO spike signal inputs.

References

Burgess,N., Shapiro,J.L. & Moore, M.A. (1991) Neural network models of list learning. *Network*, 2: 399 - 422.

French,R.M. (1992) Semi-distributed Representations and Catastrophic Forgetting in Connectionist Networks. *Connection Science, 4(3&4),* 365 - 377.

French,R.M. (1994) Dynamically Constraining Connectionist Networks to Produce Distributed, Orthogonal Representations to Reduce Catastrophic Interference. *Proceedings of the 16th Annual Cognitive Society Conference.* In press.

Greenberg,R. & Pearlman,C.A. (1974) Cutting the REM Nerve: An Approach to the Adaptive Role of REM Sleep. *Perspectives in Biology and Medicine,* 17, 513 - 521.

Grossberg,S. (1987) Competitive Learning: From Interactive Activation to Adaptive Resonance. *Cognitive Science*, 11, 23 - 63.

Carpenter,G.A. Grossberg,S. (1988) The ART of Adaptive Pattern Recognition by a Self-Organising Neural Network. *Computer*, 21(3), 77 - 88.

Gillies,A.J. (1991) *The Stability/Plasticity Dilemma in Self-organising Neural Networks*. MSc Thesis, Computer Science Department, University of Otago, New Zealand.

Hetherington,P.A. & Seidenberg,M.S. (1989) Is There “Catastrophic Interference” in Connectionist Networks? *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society,* 26 - 33. Hillsdale NJ: Lawrence Earlbaum.

Hertz, J. Krogh, A. & Palmer, R.G. (1991) *Introduction to the Theory of Neural Computation.* Redwood City CA: Addison-Wesley.

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

Lewandowsky,S. (1991) Gradual Unlearning and Catastrophic Interference: A Comparison of Distributed Architectures. In Hockley,W.E. & Lewandowsky,S. (Eds.) *Relating Theory and Data: Essays on Human Memory in Honour of Bennet B. Murdok,* 445 - 476. Hillsdale NJ: Lawrence Earlbaum.

Lewandowsky,S. & Li,S. (1994).Catastrophic Interference in Neural Networks: Causes, Solutions, and Data. Dempster,F.N. & Brainerd,C. (Eds.) *New Perspectives on Interference and Inhibition in Cognition.* Academic Press, New York. In press.

McCloskey,M. & Cohen,N.J. (1989) Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In Bower,G.H. (Ed.) *The Psychology of Learning and Motivation: Volume 23*, 109 - 164. New York: Academic Press.

McRae,K. & Hetherington,P.A. (1993) Catastrophic Interference is Eliminated in Pretrained Networks. *Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society*, 723 - 728. Hillsdale NJ: Lawrence Earlbaum.

Moody,J. (1994) Prediction Risk and Architecture Selection for Neural Networks. In Cherkassky,V., Friedman,J.H. & Wechsler,H. (Eds.) *From Statistics to Neural Networks: Theory and Pattern Recognition Applications.* NATO ASI Series F, Springer-Verlag.

Murre,J. (1992) The Effects of Pattern Presentation on Interference in Backpropagation Networks. *Proceedings of the 15th Annual Conference of the Cognitive Science Society.* Hillsdale NJ: Lawrence Earlbaum.

Nadal,J.P., Toulouse,G., Changeux,J.P. & Dehaene, S. (1986) Networks of Formal Neurons and Memory Palimpsets. *Europhysics Letters*, 1, 535 - 542.

Ratcliff,R. (1990) Connectionist Models of Recognition Memory: Constraints Imposed by Learning and Forgetting Functions. *Psychological Review,* 97(2), 285-308.

Robins,A.V. (1993) Catastrophic forgetting in Neural Networks: The Role of Rehearsal Mechanisms. *Proceedings of the First New Zealand International Two-stream Conference on Artificial Neural Networks and Expert Systems.* Los Alamitos: IEEE Computer Society Press.

Rumelhart,D.E. Hinton,G.E. & Williams,R.J. (1986) Learning Internal Representations by Error Propagation. In Rumelhart,D.E., McClelland,J.L. & the PDP Research Group (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations.* Cambridge MA: MIT Press.

Sharkey,N.E. & Sharkey,A.J.C. (1994a) Understanding Catastrophic Interference in Neural Nets. Technical Report CS-94-4, Department of Computer Science, University of Sheffield, U.K.

Sharkey,N.E. & Sharkey,A.J.C. (1994b, in press) Interference and Discrimination in Neural Net Memory. In Levy,J., Bairaktaris,D., Bullinaria,J. & Cairns,P. (Eds) *Connectionist Models of Memory and Language*, UCL Press.

White,H. (1992) *Artificial Neural Networks: Approximation & Learning Theory.* Cambridge MA: Blackwell.

Winson,J. (1990) The Meaning of Dreams. *Scientific American,* November 1990, 42 - 48.

Appendix A: Sweep rehearsal over many intervening trials

Sweep rehearsal does a very good job of maintaining the goodness of base populations over 10 to 20 intervening trials as described in the text. The question which naturally arises is over how many intervening trials can this performance be maintained?

There is no simple answer, but rather a range of factors to be considered. In the case of heteroassociative learning the goodness of the previously learned items deteriorates – perhaps more gradually than one might expect – as more and more intervening items are added. In the case of autoassociative learning, however, the goodness of the previously learned items remains relatively high as a large number of items are added. We found little deterioration in performance in trials using up to 70 intervening items. However the success of the network in the autoassociative case is far from unqualified. To a large extent an autoassociative network which has learned a number of patterns is no longer recalling a specific population of learned vectors, but is rather reproducing or "passing through" any input that it is given. The distinction between these two conditions is that in the case of genuine learning a network would be able to distinguish between the learned population and any new test population of novel inputs ("recognise" the learned population), whereas in the case of learning to pass through any input it would not. The incorporatation of the second kind of learning by autoassociative networks, tending to reproduce any input that it is presented with, was first drawn to my attention by Dr Noel Sharkey (personal communication), who called the condition "catastrophic remembering".

In the remainder of this appendix we present these results in more detail. The ability to discriminate between a previously learned / old population and completely new / novel items can be quantified using a measure of discriminability as discussed extensively by Ratcliff. The discriminability d' is the difference in the means of the goodness values for the old and the novel populations, divided by the standard deviation of the goodness values for the novel population. Note that we use the terms "old" and "novel" in the sense described above for the remainder of the appendix; old patterns are previously learned patterns including the base population and any learned intervening items, novel patterns are a newly generated test population (consisting of 70 randomly generated items – without duplication of old items – in our simulations).

The behaviour of sweep rehearsal over a large number of intervening trials is illustrated in the following figures, which also enable us to examine the behaviour of the *d'* measure and its components. We call the graphs shown in these figures *discriminability graphs*. Simulations are based on the same architecture and methods as described in the text for sweep rehearsal. The network is trained on a base population of 20 items as before, but 70 intervening trials – far more than in previous simulations – are then added. After each intervening trial the network is tested with both the old population and a novel population so that the measures relevant to the discriminablity of the old and novel populations can be calculated and plotted for that population size. What we have so far in this paper called the goodness of the population is thus on these discriminability graphs represented by the measure “old population mean”. Strictly speaking this is the average of each individual learned item’s goodness after training is complete – a given serial position curve such as those shown in Figures 8 and 10 could be averaged to produce a single point (the old population mean at “X axis” position 20) for a discriminability graph.

Figure A1 shows the discriminability graph for the heteroassociative learning condition. As described above the “goodness of the population” (the old population mean measure) declines slowly, but steadily, as the population size increases. There is also a steady fall in the discriminability measure. This is a result of the decline in the old population mean, as other components remain roughly constant.

Figure A2 shows the discriminability graph for the autoassociative learning condition. In this case the “goodness of the population” (the old population mean measure) remains remarkably high – after 70 intervening items the goodness is still approximately 0.9! As noted above, however, this apparent success masks an underlying problem. The network is not necessarily learning the population in any useful sense. Although it can reproduce learned items with a high degree of accuracy it has difficulty in distinguishing these “learned” items from completely novel inputs, as shown by the initially low and gradually declining discriminability. In this case the low discriminability is mediated not by a fall in old population mean, but by a high novel item mean component. In retrospect this is not a surprising result. In the autoassociative paradigm the network can achieve successful performance not only by learning individual items, but also by learning to simply reproduce or “pass through” any item that it is given. Any kind of autoassociative learning (including ordinary training, sweep rehearsal, and sweep pseudorehearsal) will suffer from this effect as a number of patterns are learned. The extent to which the network has adopted such a strategy in this case is reflected in the novel item mean in Figure A2 – after learning the base population a high novel item mean is already well established. This is a subtle effect as the goodness of each old (previously learned) item remains high, and nothing in the networks “visible behaviour” indicates that the underlying discriminability of the items is poor (unless it is explicitly tested on novel items).

In summary, we have explored the behaviour of sweep rehearsal training over a large number of intervening trials. Performance declines gradually. In the heteroassociative case this is reflected in an obvious decline in goodness, and on testing a fall in discriminability is also apparent. In the autoassociative case the goodness of the population remains high. However testing shows that there is a fall in discriminability in this case also – as a by-product of autoassociative learning the network is tending to reproduce not only the learned population, but any input it is given. Given the effectiveness of sweep rehearsal, we suggest that these limitations are manifestations of fundamental upper limits on the performance of rehearsal mechanisms in general. The greater the number of items stored the more likely it is that representational overlap (see Section 6.2) will cause these items to interact with each other. While an effective rehearsal mechanism will allow many additional items to be successfully added, the constraints imposed by representational overlap in a finite architecture will eventually manifest themselves in falling goodness and / or discriminability.

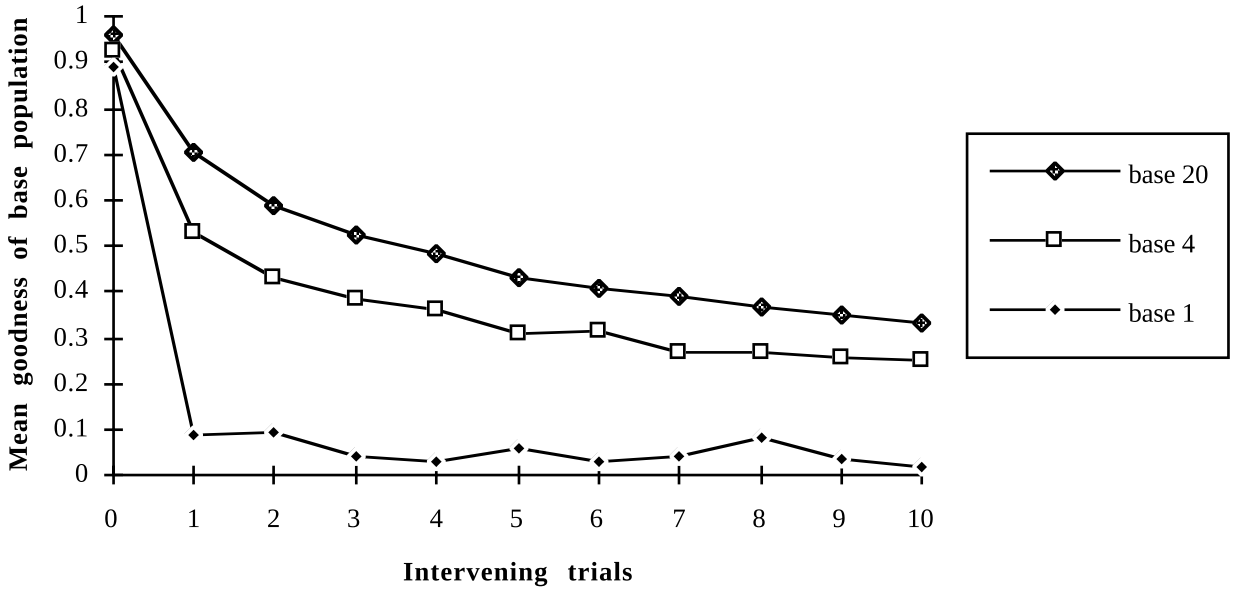

Figure 1: Fall in base population goodness as 10 intervening trial items are added. Base populations of size 1, 4, and 20 are shown.

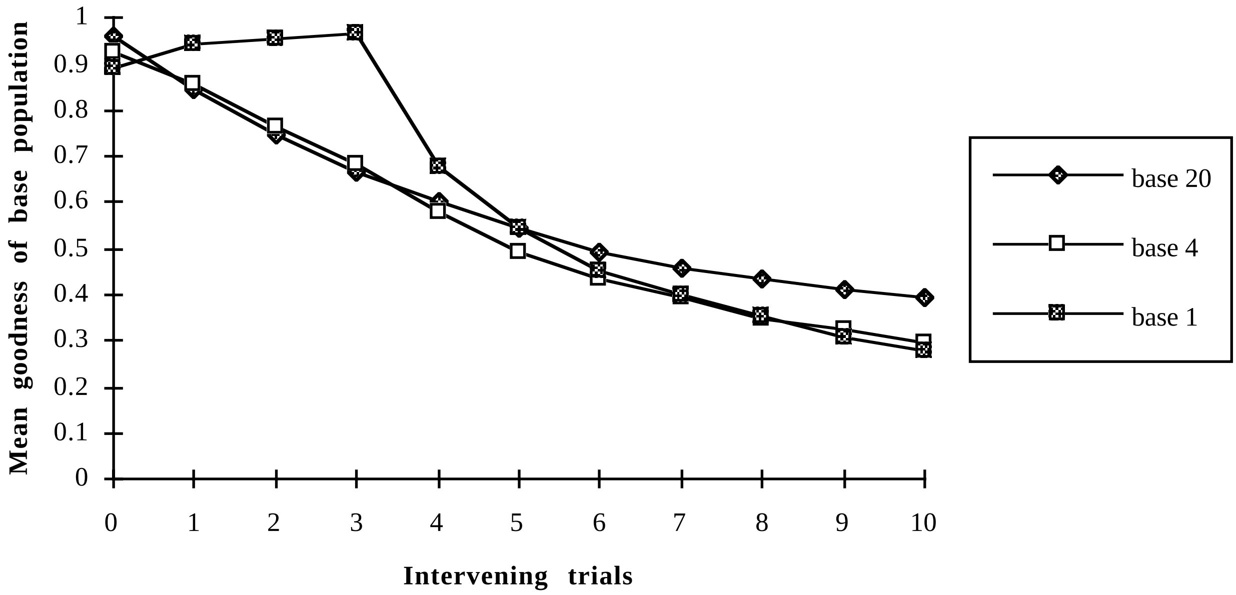

Figure 2: The effect of recency rehearsal on base population goodness as 10 intervening trial items are added. Base populations of size 1, 4, and 20 are shown.

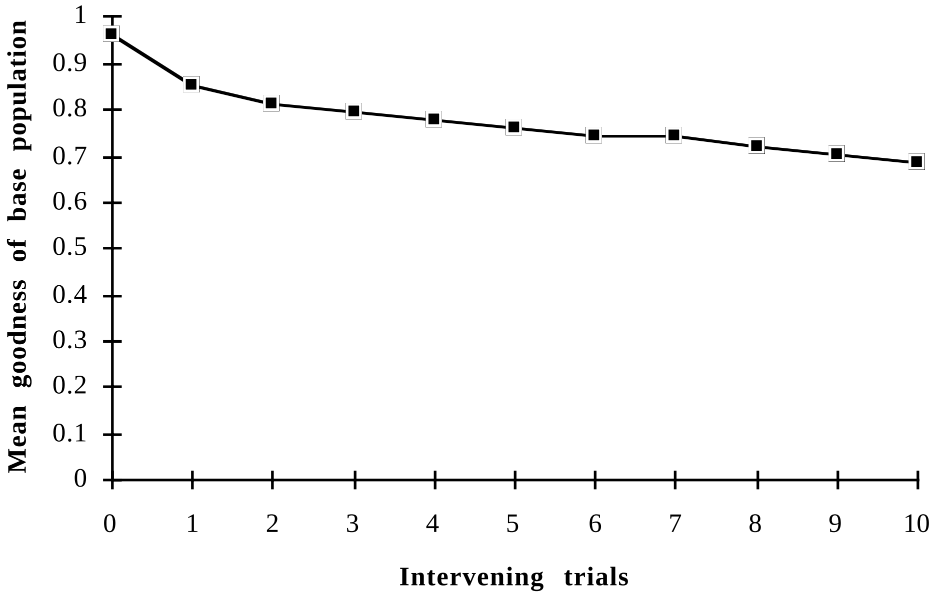

Figure 3: The effect of random rehearsal on base population goodness as 10 intervening trial items are added. The base population consists of 20 items.

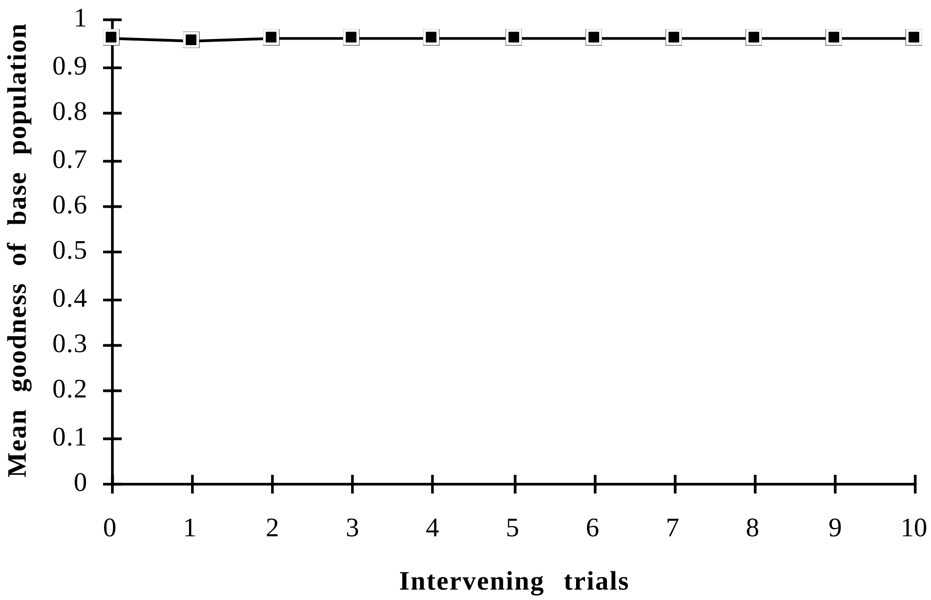

Figure 4: The effect of sweep rehearsal on base population goodness as 10 intervening trial items are added. The base population consists of 20 items.

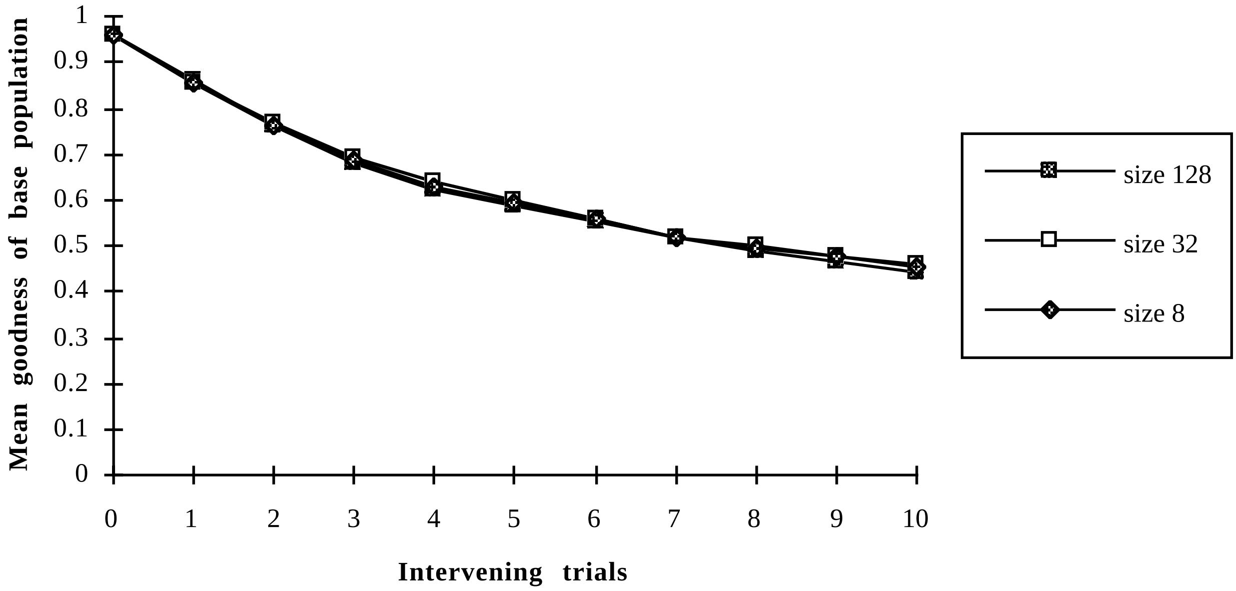

Figure 5: The effect of random pseudorehearsal on base population goodness as 10 intervening trial items are added. The base population consists of 20 items, and pseudopopulations (used in rehearsal) of size 8, 32, and 128 are shown.

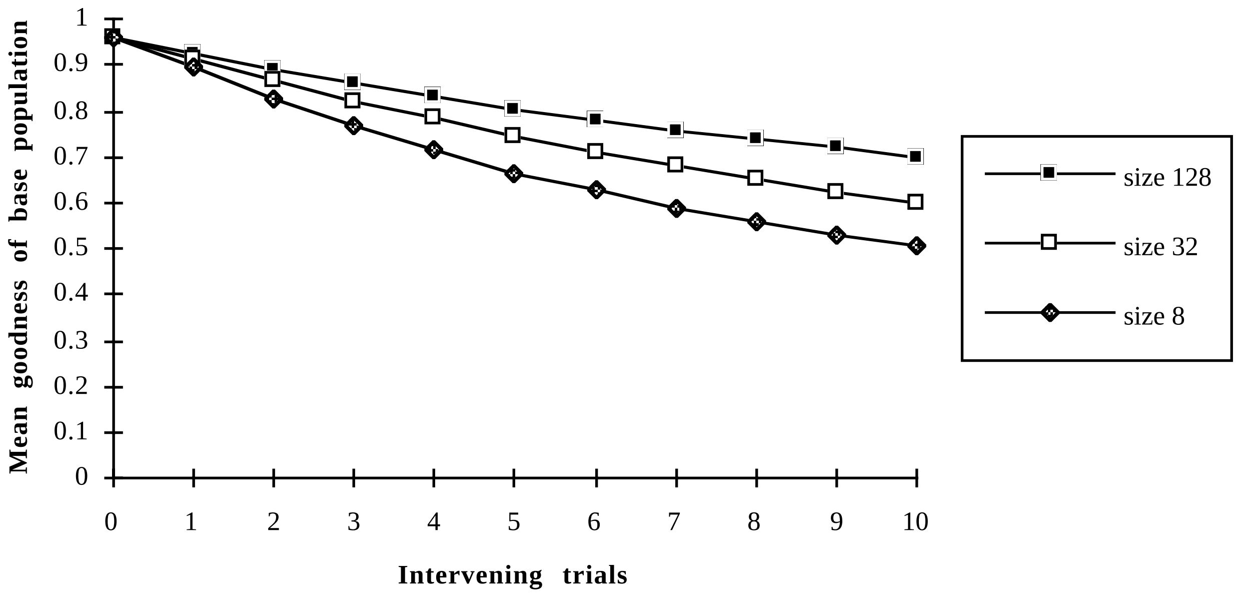

Figure 6: The effect of sweep pseudorehearsal on base population goodness as 10 intervening trial items are added. The base population consists of 20 items, and pseudopopulations (used in rehearsal) of size 8, 32, and 128 are shown.

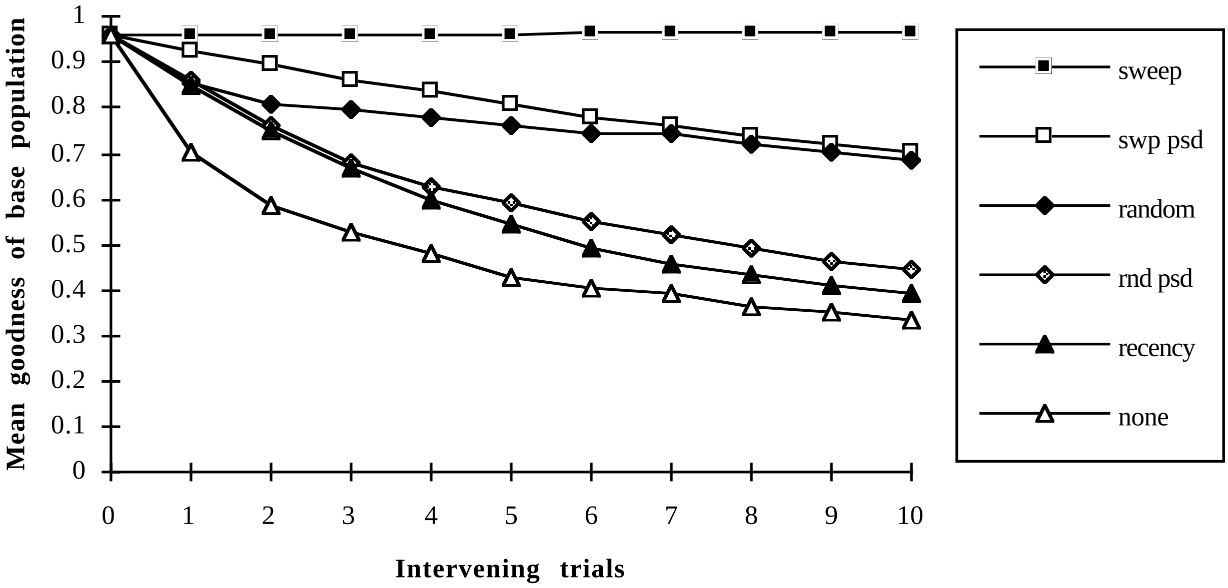

Figure 7: Summary of results presented so far for base population goodness over 10 intervening trials. Heteroassociative population. Conditions shown are sweep rehearsal, sweep pseudorehearsal, random rehearsal, random pseudorehearsal, recency rehearsal, and no rehearsal.

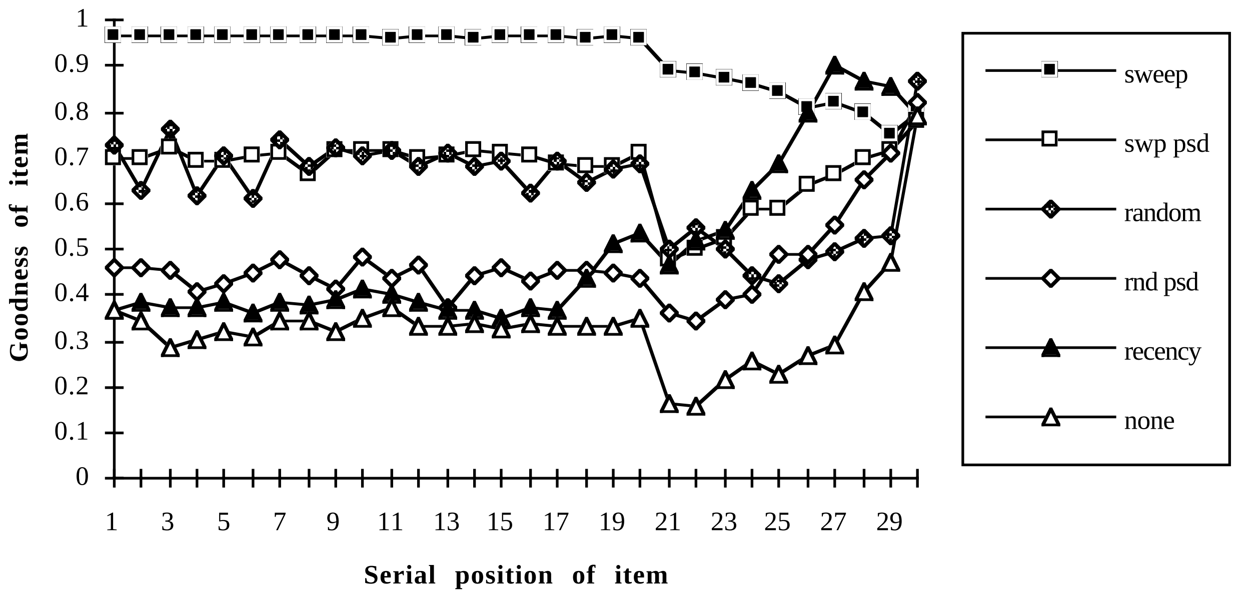

Figure 8: Summary of goodness of all items by serial position after training is complete. Heteroassociative population. Conditions shown are sweep rehearsal, sweep pseudorehearsal, random rehearsal, random pseudorehearsal, recency rehearsal, and no rehearsal.

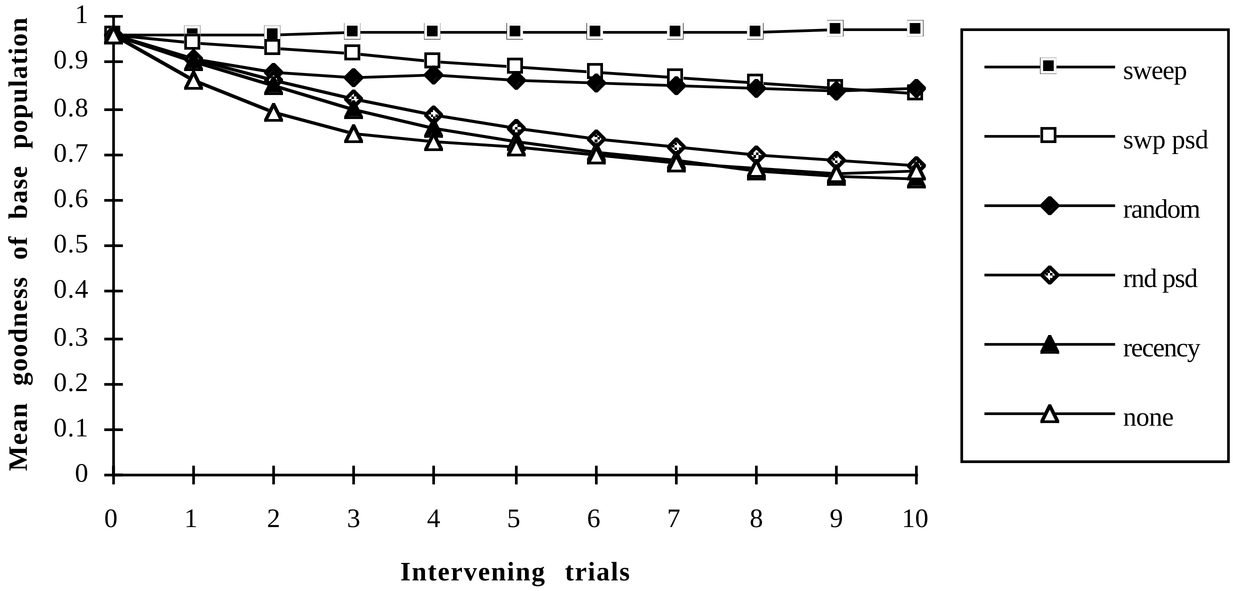

Figure 9: Base population goodness summary using an autoassociative population. Conditions shown are sweep rehearsal, sweep pseudorehearsal, random rehearsal, random pseudorehearsal, recency rehearsal, and no rehearsal.

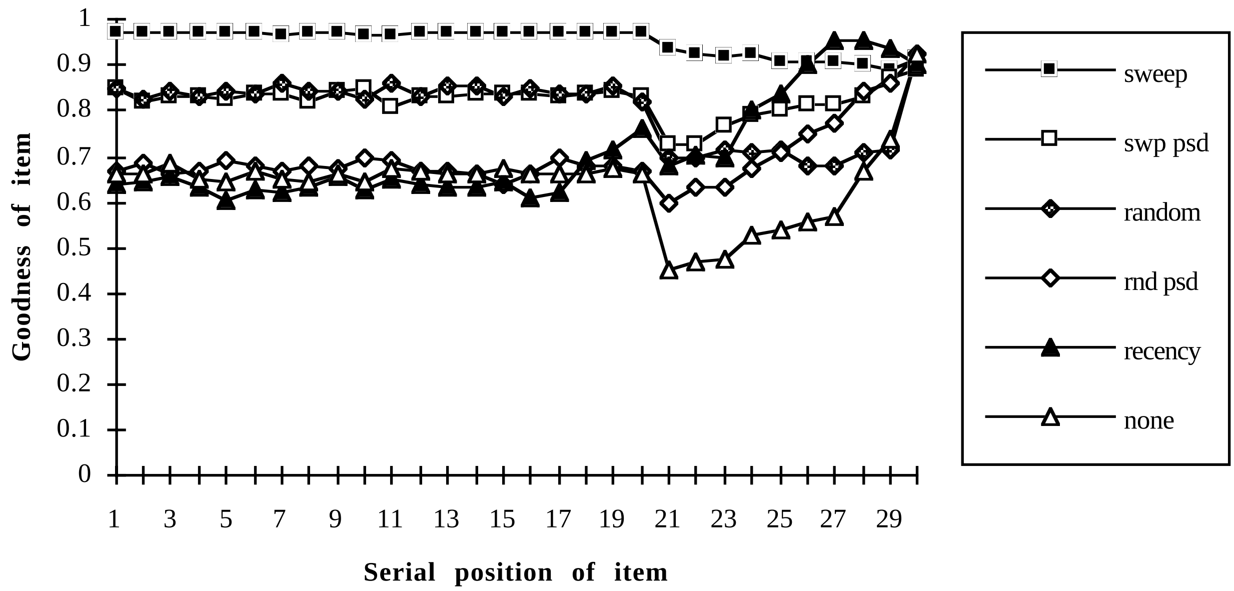

Figure 10: Serial position summary using an autoassociative population. Conditions shown are sweep rehearsal, sweep pseudorehearsal, random rehearsal, random pseudorehearsal, recency rehearsal, and no rehearsal.

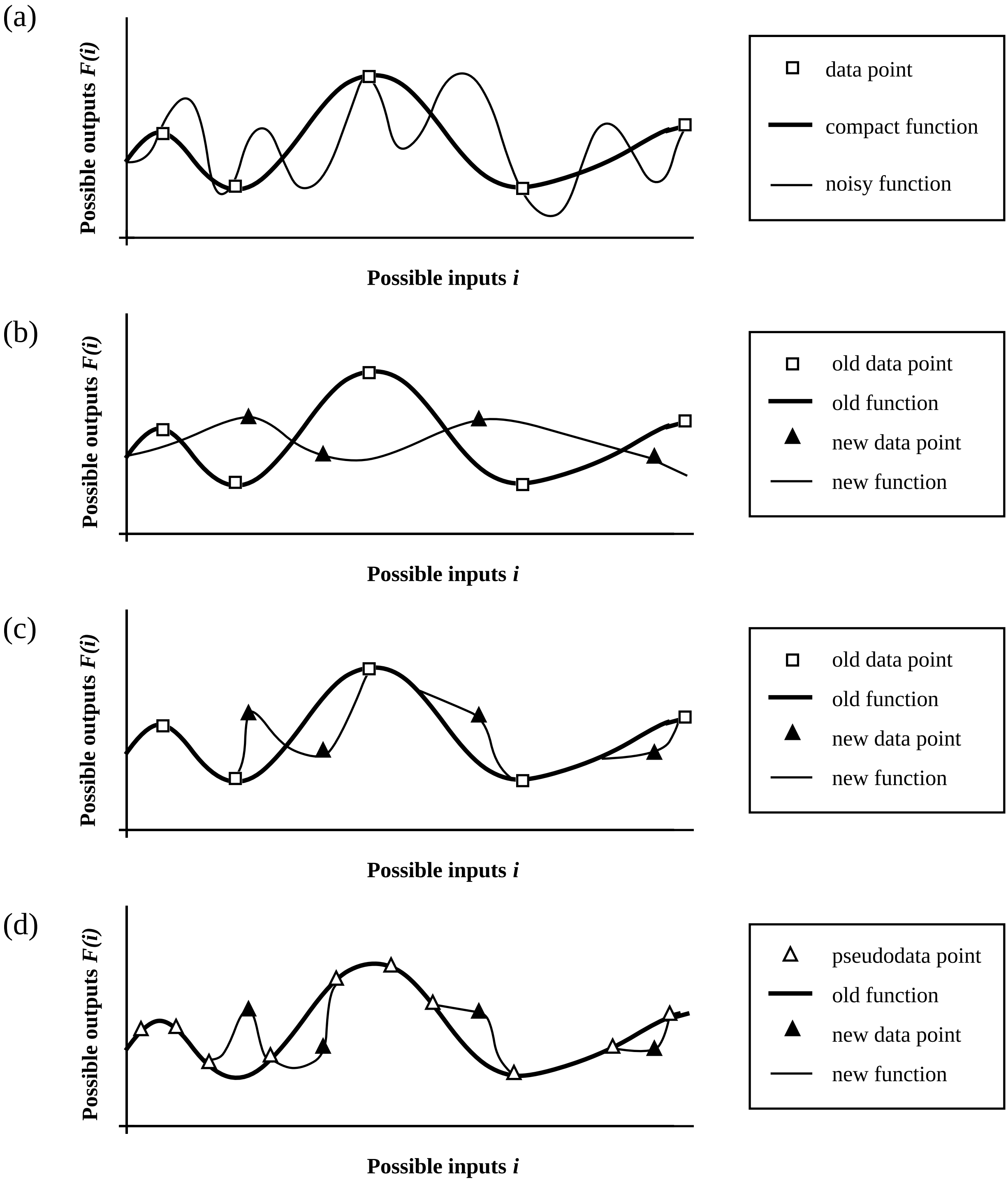

Figure 11: (a) Interpreting learning in a neural net as function approximation. Learning fits a function to the training population data points, a range of functions are possible including “compact” and “noisy” functions. (b) New intervening item data points are learned without rehearsal, the new learned function may not be similar to the old function describing the original population. (c) New intervening item data points are learned with rehearsal, the new learned function will preserve much of the shape of the old function. (d) New intervening item data points are learned with pseudorehearsal using pseudodata points, the new learned function will preserve much of the shape of the old function.

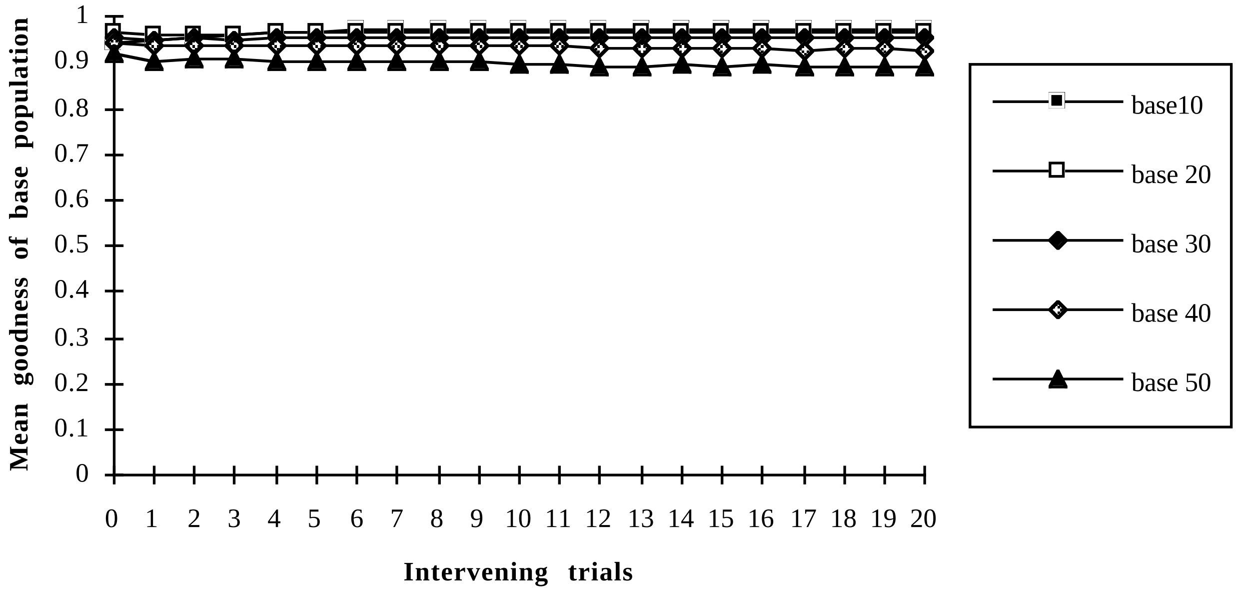

Figure 12: Comparison of sweep rehearsal performance for base populations of sizes 10 to 50 over 20 intervening trials. Heteroassociative population.

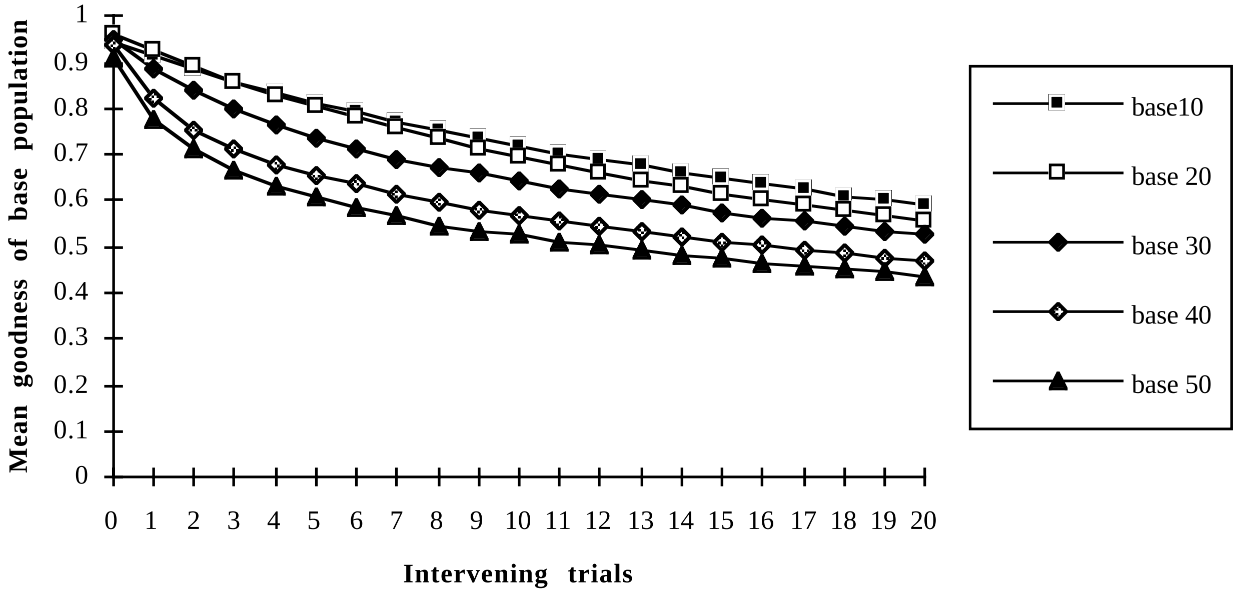

Figure 13: Comparison of sweep pseudorehearsal performance for base populations of sizes 10 to 50 over 20 intervening trials. Heteroassociative population.

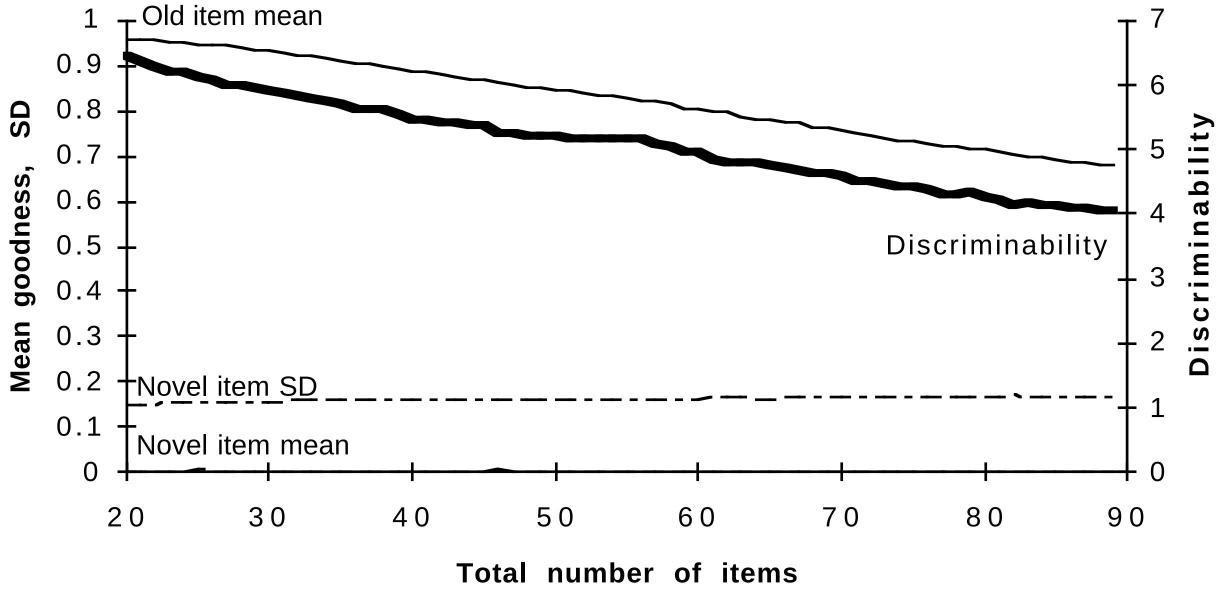

Figure A1: Discriminability graph for heteroassociative population learning. Shows the discriminability measure (d') and its components for populations with total numbers of items from 20 (base population alone) to 90 (base population plus 70 intervening trials).

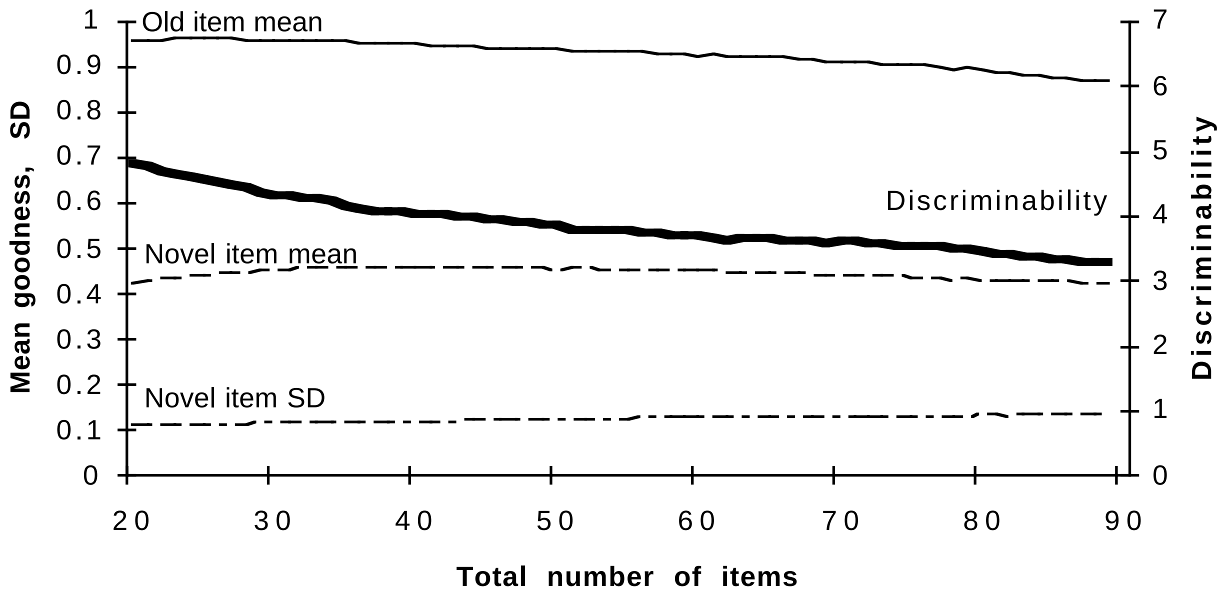

Figure A2: Discriminability graph for autoassociative population learning. Shows the discriminability measure (d') and its components for populations with total numbers of items from 20 (base population alone) to 90 (base population plus 70 intervening trials).