

Control of consolidation in neural networks: avoiding runaway effects

Martijn Meeter

Department of Psychology, University of Amsterdam, The Netherlands

email: m@meeter.nl

Abstract. Consolidation has been implemented in two ways: as straight rehearsal of patterns or as pseudorehearsal, in which pseudoitems are created by sampling attractors or input–output combinations from the network. Although both implementations have been investigated by several authors, few have explored how it is decided which pattern or pseudoitem is consolidated. Controlling consolidation is not trivial, as it is susceptible to a corruption. In runaway consolidation, one or two patterns monopolize all consolidation resources and come to dominate the entire network. Runaway consolidation is analysed, and three solutions are explored. Suppressing transmission in the connections in which consolidation takes place is shown to work best. Placing bounds on connections or unlearning attractors also alleviates runaway consolidation, though less effectively so.

Keywords: consolidation, rehearsal, pseudorehearsal, catastrophic interference.

1. Introduction

Consolidation, as a concept, has a century-old history (Lechner *et al.* 1999). It was originally proposed as an explanation for retroactive interference (Muller and Pilzecker 1900). Although interference theory soon came to rely on other constructs (McGeoch 1932), consolidation fanned out to explain a plethora of other phenomena. It is now used as a label for very different processes that operate on widely varying time-scales (Squire and Alvarez 1995). Even within the limited domain of the neural networks literature, consolidation has been the answer to several unrelated problems, though sometimes under different names (e.g. self-refreshing or synaptic reentry reinforcement). It is perhaps most often suggested as an explanation for the Ribot curve (Alvarez and Squire 1994, McClelland *et al.* 1995, Murre 1996, Meeter and Murre, submitted), but the same processes have also been proposed as an answer to catastrophic interference (McClelland *et al.* 1995, Robins 1996, Ans and Rousset 1997, Robins and McCallum 1998, 1999), and Murre *et al.* (submitted) have used a method of self-repair akin to consolidation to simulate recovery from mild brain damage (Robertson and Murre 1999).

Several researchers have carried out network simulations implementing a form of consolidation. Nevertheless, not much attention has been paid to how a consolidation process is to be controlled from within a model. In one implementation in which a neural network itself selected the pattern to consolidate, it was observed that under certain

Correspondence should be addressed to M Meeter, Centre for Molecular and Behavioural Neuroscience, 197 University Avenue, Newark, NJ 07102, USA.

conditions the process was susceptible to a curious corruption, runaway consolidation (Meeter and Murre, submitted). One pattern would, by chance, benefit more from consolidation than others, and through that fact start monopolizing all consolidation resources. In the end, only that pattern would still be an attractor in the network.

This paper will argue that runaway consolidation is a rather general problem of consolidation if Hebbian learning is used, one that easily appears and is not tied to a particular implementation. It will start with a short review of the roles that consolidation has played in computational studies, and how it has been implemented. Then, the problem of runaway consolidation will be introduced with two simple simulations. In the third part of the paper, three possible solutions will be proposed and explored.

1.1. *Catastrophic interference*

One problem for which consolidation has been suggested as the solution is catastrophic interference (McClelland *et al.* 1995). Catastrophic interference occurs in several connectionist paradigms—notably in networks using the ‘back-propagation’ algorithm—when a network is to acquire sequentially two or more pattern sets. In a typical demonstration, one pattern set is learned to criterion, and the network is then trained on a second set. This procedure results not in mediocre performance on both sets, but in good performance on the second set, and a total erasure of all patterns in the first. The second set interferes with the first one with catastrophic consequences (McCloskey and Cohen 1989, Ratcliff 1990, French 1999).

Many solutions to catastrophic interference have involved *interleaved learning*, basically mixing in, during training on the second pattern set, patterns of the first set. This protects the first set, usually referred to as the base patterns, from interference. Interleaved learning has been implemented in two ways. The first, straight rehearsal of patterns, refers to relearning either the whole or part of the base population during each training session with a new pattern (McClelland *et al.* 1995, Murre 1992, Ratcliff 1990, Robins 1995). Although this circumvents the effects of learning new patterns, Robins and McCallum (1999) rightfully point out that it presupposes that the whole base population is stored in a separate system reinstating patterns for rehearsal. This set-up would make the original network a little superfluous: one could just as well use the reinstating system as the memory (although the two memories might perform different functions, see McClelland *et al.* 1995).

The second technique, ‘pseudorehearsal’, does not suffer from this drawback. Instead of rehearsing with the base population, a population of ‘pseudoitems’ is created from the network and used for rehearsal purposes. These pseudoitems are patterns generated from the network after the network has been trained on the base population. As in rehearsal, these pseudoitems are then inserted into the learning set during each training session with new patterns. In simulations with a Hopfield network, pseudoitems are generated by letting the network relax from a random initial state to an attractor (Robins and McCallum 1998). This attractor is then inserted in the pseudoitem population. In simulations with back-propagation models, they consist of random input patterns and the outputs generated by these patterns (Robins 1995, 1996). These two methods can also be merged, as was shown by Ans and Rousset (1997, 2000).

Although not as effective as plain rehearsal, pseudorehearsal does protect base patterns from interference by the new pattern set. This is because pseudoitems can be seen as samples from the network, identifying the function that the network is calculating—or, in the Hopfield framework, the attractor landscape in the network. When a pseudoitem population is created after the network has incorporated the training set, then the pseudoitems sample the function (or attractor landscape) that the base population amounted to.

Training the network on a suitably large sample of pseudoitems therefore implies training it on the same function (or attractor landscape) as with the base population (Robins and McCallum 1999).

1.2. Cortico–hippocampal interactions

Both rehearsal and pseudorehearsal have cropped up in a very different context, that of cortico–hippocampal interactions, memory consolidation and retrograde amnesia. After damage to the hippocampal memory system, patients tend to lose more of their recent memories than their distant memories (Ribot 1881, Squire 1992), a pattern referred to as the Ribot gradient. This can be explained by assuming that memories are initially retrieved via a hippocampal memory system. Through a process of consolidation, memories gradually become stored in the neocortex, making them independent of the hippocampal system (Squire and Alvarez 1995, Squire *et al.* 1984). If the hippocampal system is damaged, recent memories are lost because they are still dependent on that system. Old memories, however, have already been stored in the neocortex through consolidation and are thus spared.

Three computational models have simulated such cortico–hippocampal interactions (Alvarez and Squire 1994, McClelland *et al.* 1995, Murre 1996, Meeter and Murre, *in press*). All three share the assumption that representations stored in the hippocampal system form the basis of representations gradually built up in a neocortical memory system during consolidation.

McClelland *et al.* (1995) did not explicitly simulate a hippocampal system; consolidation consisted of relearning old patterns according to a probability distribution that was assumed to incorporate the behaviour of the hippocampal system. This comes down to rehearsal. Both other models of retrograde amnesia implemented forms of pseudorehearsal. Meeter and Murre (*in press*) simulated the neocortical memory system as a large layer in which only weak connections could be made between nodes belonging to one pattern. These nodes were indirectly bound via hippocampal nodes with strong connections to all neocortical nodes in the pattern. Consolidation was simulated by letting the model, from an initial random state, relax into an attractor, and then updating the weights with a Hebbian learning rule. In the much smaller model of Alvarez and Squire (1994), patterns consisted of four neocortical nodes connected to one hippocampal system node (they used the name *medial temporal lobe system*). Consolidation consisted of activating a random hippocampal node and letting the model cycle for three iterations. Weights were updated while the node activated its associated neocortical pattern. In both models, the connections between neocortical nodes built up during consolidation eventually allowed the patterns to be retrieved without the support of the hippocampal system.

Although their goals were different (modelling brain repair and synaptic re-entry, respectively), both Murre *et al.* (*submitted*) and Wittenberg *et al.* (2002) proposed forms of pseudorehearsal akin to Alvarez and Squire's (1994) and Meeter and Murre's (*in press*) tack on consolidation. They stored a number of patterns in a Hopfield-type network, let their model find an attractor and adapted the weights in the network to this attractor.

1.3. Which pattern is consolidated?

An important question surrounding consolidation is what determines which pattern is to be consolidated. Pattern choice has often been relegated to structures not explicitly modelled (McClelland *et al.* 1995, Robins 1996, Robins and McCallum 1998). For example, McClelland *et al.* (1995) did not simulate a system that restores an old pattern for an

Wittenberg *et al.* 2002, Meeter and Murre, submitted, Murre *et al.*, submitted). This method has functioned reasonably well: usually, the attractors found by the network were stored patterns. There is, however, a danger with this method: runaway consolidation (Meeter and Murre, submitted).

2. Runaway consolidation

2.1. Demonstration with a Hopfield net

The problem of runaway consolidation can be illustrated with a simple set of simulations, loosely modelled after those of Robins and McCallum (1998). As in that paper, a base set was stored in a network, and the goal was to protect the base set against interference by later patterns. In contrast to Robbins and McCallum, a simple Hopfield network was used (Hopfield 1982; Robbins and McCallum used asymmetric weights and an error-correcting learning rule). The size of the Hopfield net was 80 nodes. Patterns used in the simulation consisted of a random 40 nodes with value 1, and 40 with value of -1 . By updating the weights in the network according to a simple Hebbian rule, one can transform such patterns into attractors of the network (Hopfield 1982). Fifteen patterns were stored in the network in this way, which is above the normal capacity of a network of 80 nodes. The first five were designated as the base set, the other 10 as interfering items. The network was tested after it had acquired the base set of five patterns, and again after storage of each new interference pattern. In these tests the cue consisted of 50% of the target pattern.

Figure 2 shows the performance of the network, based on 30 replications. After storing the base set, all five patterns of that set could be retrieved perfectly from a 50% cue (see figure 2(a)). With each additional stored pattern, however, performance dropped, until the mean Hamming distance at retrieval was around 20 for both patterns in the base set and the subsequent patterns (the order in which patterns are learned is unimportant in basic Hopfield networks).

A second simulation was done with the pseudorehearsal method of Robins and McCallum (1998), again with 30 replications. After the five patterns in the base population had been learned for each 50 learning trials (equivalent to updating the weights once with a learning parameter of 50), a pseudoitem population of 50 attractors was created. One such attractor was found by inserting a random pattern in the network and then letting it iterate for 2000 steps. It was assumed that after those steps a stable attractor had been reached, and this attractor was then entered in the pseudoitem population. As in the simulations of Robins and McCallum (1998), there was no check on whether items belonged to the base population or not. Unlike in those simulations, there was also no check on whether pseudoitems were duplicates of one another (in such a simple network, finding 50 independent attractors could take a prohibitively long time). The number of 50 pseudopatterns was chosen to be large enough to limit the likelihood of one pattern being repeated often in the pseudopopulation. A higher number of pseudoitems did not improve performance.

Each new pattern received 50 learning trials as well. Intermingled with those, the 50 pseudoitems were learned for one learning trial each. As can be seen in figure 2(b), this reduced, though not eliminated, the number of errors when patterns in the base set were tested. New patterns, however, suffered from this procedure: after 15 patterns had been learned, performance on the new patterns was barely above its minimum value of 40 (Hamming distances lower than 40 were subtracted from 80 to give the distance to the mirror image of the pattern). This result was also found by Robins and McCallum (1998). Pseudorehearsal is thus a good strategy for reducing interference by subsequent items, though at the cost of a reduced encoding of those subsequent items.

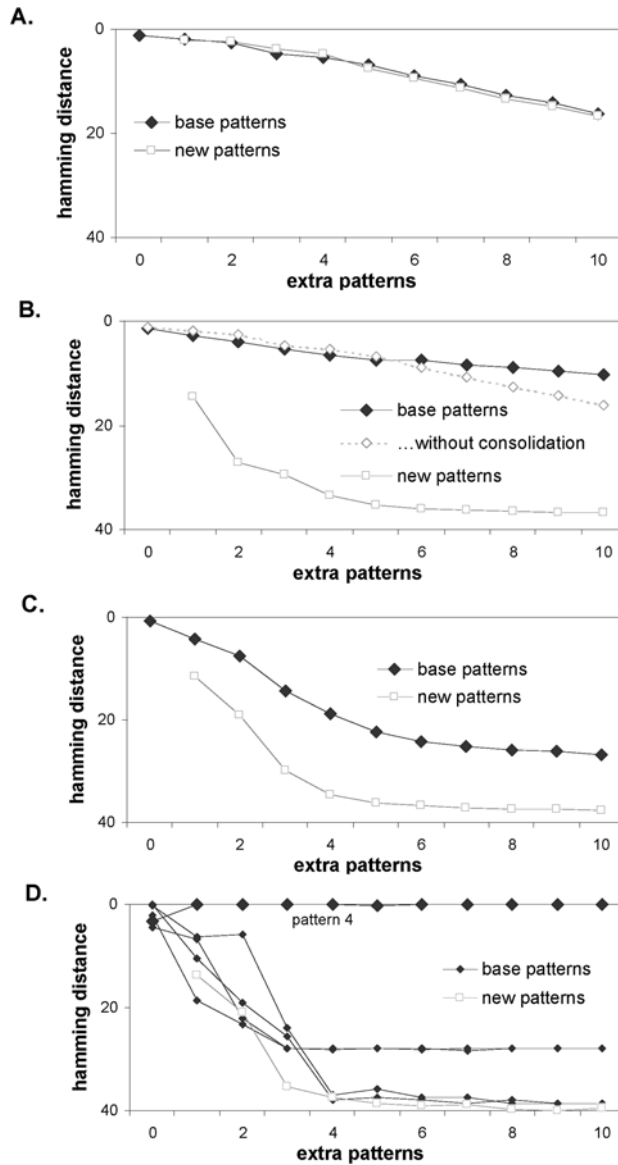


Figure 2. Demonstration of runaway consolidation in a Hopfield network. After a base pattern set of five patterns is stored the network is tested, and again as more and more interfering patterns are stored. The x-axis plots the number of interfering patterns stored, the y-axis the mean Hamming distance at test for both base patterns and interfering patterns (under ‘new patterns’) (a) Simulation without consolidation. (b) With consolidation through a pseudo-item population, base patterns are better resistant to interference from new patterns. Interfering patterns suffer and are not stored very well. To ease comparison, the base patterns line from (a) is also plotted (‘without consolidation’). (c) With pseudoitems not taken from a population but sampled real-time from the network, performance suffers because of runaway consolidation. (d) A single replication of the simulation in (c) shows that one base pattern (pattern 4) is retained very well, while other base patterns are lost from the network. Pattern 4 monopolizes all consolidation trials, it ‘runs away’ with the consolidation process.

Pseudorehearsal as practised above assumes that a training system holds the population of pseudoitems for the duration of the simulation (Robins 1997). If no such system is assumed, pseudoitems have to be created during the phase of pseudorehearsal. In a third simulation, this was done. As in the previous ones, the model first learned the base set and then the interfering patterns. Intermingled with the 50 learning trials for the new items were, again, 50 pseudorehearsal trials. Unlike in the previous simulation, however, the pseudoitem to be rehearsed was created anew at each pseudorehearsal trial by letting the model find an attractor as above. This resulted in a dramatic worsening of performance (see figure 2(c)); performance was, under this regime, even worse than performance without consolidation (figure 2(a)).

A look at a single replication of the third simulation (figure 2(d)) shows what is going on. One pattern in the base pattern set, pattern 4 in this replication, is still retrieved perfectly. Meanwhile, all other patterns in the base set, as well as new patterns, are retrieved only at chance level. There is a simple reason for these results: nearly all pseudorehearsal trials resulted in the consolidation of an attractor equal or similar to one pattern. This pattern then remained retrievable, while all other patterns in the base set were lost. Pseudorehearsal thus performs surprisingly badly once a whole pseudoitem population is no longer stored. Another example of this tendency was given by Wittenberg *et al.* (2002), who stored five patterns in a Hopfield network but found that consolidation left only one of these intact after a few trials.

The deeper cause is that consolidating a pattern increases the likelihood that a pattern is consolidated in the next run. Consolidated patterns are thus consolidated more often, which in turn assures that they will receive an even bigger share of future consolidation trials. This quickly leads to what Meeter and Murre (submitted) have called ‘runaway consolidation’ one or two patterns ‘run away’ with the consolidation process and monopolize all consolidation resources.

2.2. *The TraceLink model*

Runaway consolidation does not automatically disappear with the addition of a training system to a model. A ‘daytime’ functioning in which the target memory is strong enough to be relevant, must be married with a ‘night-time’ functioning in which the training system is strong enough to determine which memory is consolidated into the target memory. A training system that is too extensive will take over the function of memory, thereby making the target memory superfluous. A target memory that is too strong will dominate consolidation and therewith lead to runaway consolidation.

One model in which the distinction between a training system and a target memory has been implemented is TraceLink, one of the computational models that has been used to simulate remote memory and amnesia (Murre 1996, Meeter and Murre, in press, submitted). The TraceLink model illustrates with connectionist simulations how a process of consolidating memories in the neocortex with help of the hippocampus may explain many characteristics of amnesia, including gradients in retrograde amnesia, shrinkage of retrograde amnesia during recovery and transient global amnesia (Meeter and Murre, submitted).

TraceLink consists of two layers: a trace system modelled as a layer of 200 nodes and a link system modelled as a layer of 42 nodes. The trace system corresponds to the neocortex, the link system to a medial temporal lobe temporary storage system. Both layers have internal connections and are connected with one another. Between every two nodes an excitatory connection can be formed.

Both layers have binary stochastic nodes. The likelihood of firing of these nodes depends on the balance between the excitatory input from other nodes and inhibition.

The excitatory input to a node is the weighted sum of the activation of all nodes connected to it. Inhibition is a sum that is constantly fine-tuned so as to keep the average number of active cells in a layer as close to a preset target number (k) as possible. This number is set separately for every layer, and consequently inhibition is regulated separately in every layer. The weights of connections are modifiable with a variant of Hebb's rule (Hebb 1949) that allows for weight increases as well as decreases: weights are increased with the learning rate μ if both pre- and postsynaptic nodes fire, and are decreased with a value $0.75 * \mu$ if the postsynaptic node fires but the presynaptic one does not. Weights are clipped to the interval $[0, 1]$. For more details, see Meeter and Murre (in press).

Learning is not equally fast for all connections. The learning rate (μ) is much lower for the within-trace connections ($\mu = 0.06$) than for the connections within the link layer, or between the link layer and the trace layer (both $\mu = 0.4$). This models the fact that the hippocampus is unusually plastic (Lopes da Silva *et al.* 1990) and that the regions of the link system have a higher connectivity than the neocortex (Treves and Rolls 1994).

A pattern in the learning set consists of 10 random trace nodes and 7 random link nodes, which are activated when a pattern is presented. The number of nodes in the two layers that belong to a pattern is equal to the target number of active nodes in equilibrium for that layer (e.g. $k = 10$ in the trace layer). As patterns are chosen randomly, every two patterns share on average a k/m proportion of their nodes in a layer, where m is the number of nodes in the layer.

For acquisition, a pattern is presented and then learned for one iteration with the learning rates given above. Consolidation is implemented by letting the model cycle for 150 iterations and storing the attractor that surfaces. Only trace-trace connections are modified in consolidation, connections within link and between trace and link are not changed. Retrieval of a pattern is measured by activating a cue of 30% of the trace portion of the pattern (three random trace pattern nodes) and letting the model cycle for 80 iterations. The proportion of uncued trace pattern nodes that are active after these 80 iterations is taken as the measure of retrieval. This can thus vary between zero and one.

In a basic TraceLink simulation of remote memory, 15 patterns are stored. After acquisition of each new pattern, there is a consolidation period in which three autonomously surfacing attractors are strengthened. Only trace-trace connections are modified in consolidation, connections within link and between trace and link are not changed. When the 15 patterns are stored, retrieval for each pattern is tested, both in an intact condition and in a lesioned condition in which the Link system is deactivated. In this second condition simulating amnesia through medial temporal lobe damage (which will not be discussed further here), recent patterns are lost, while old consolidated patterns are still retrievable.

2.3. Runaway consolidation in the TraceLink model

The consolidation learning parameter is usually set to a value where the three consolidation trials in one consolidation period cause as much weight change in the trace layer as initial acquisition of one pattern does (value 0.06). To illustrate runaway consolidation, we repeated the basic simulation outlined above (without amnesia condition) with different strengths of consolidation. The black line in figure 3(a), based as the other lines on 100 replications, shows performance in the intact model without consolidation. Recent patterns (to the right) are retrieved fairly well, while older patterns (to the left), through overwriting and interference, are retrieved progressively worse. With consolidation of standard strength of initial learning, remote patterns are less vulnerable to interference. This occurs without large costs to the retrievability of recent patterns. However, if consolidation is too strong compared with acquisition learning (i.e. total weight change

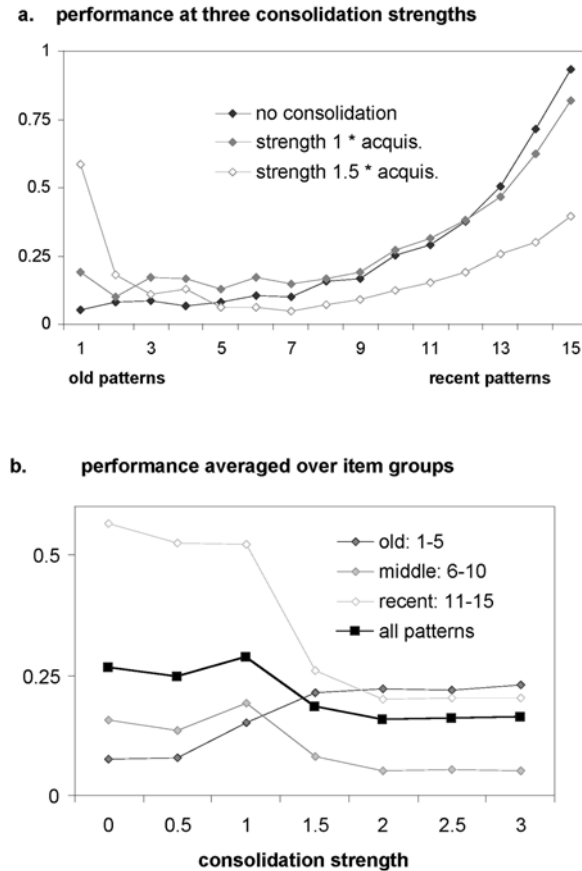


Figure 3. Runaway consolidation in TraceLink fifteen patterns are stored, with acquisition of new patterns intermingled with consolidation trials of a strength varied across simulations. Consolidation strength is given as multiples of the trace acquisition learning parameter. (a) Mean retrieval of the 15 individual patterns at three consolidation strengths: no consolidation, consolidation with a strength equal to one acquisition trial and consolidation of 150% that strength. The latter line shows the effects of runaway consolidation (b) Mean retrieval at different consolidation strengths for all patterns, for remote patterns (first five to be acquired), middle patterns (patterns 6–10) and recent patterns (patterns 11–15). Though with moderate consolidation strength mean pattern retrieval increases, with higher strengths (>1 time acquisition strength) it decreases again. Remote patterns still benefit from increases in consolidation strength, but both middle and recent patterns suffer.

during consolidation is greater than that during learning), consolidated patterns tend to ‘run away’ in the same way as illustrated above with the Hopfield model. One pattern, usually the first, becomes so strong through consolidation that it monopolizes all consolidation resources and is the sole strengthened pattern. The difference between consolidation and initial learning does not have to be very large for this to occur. The grey line with open markers in figure 3(a) shows runaway consolidation, with consolidation weight change being just 50% higher than initial learning. In this setting, only pattern 1 could be retrieved very well, as this pattern had received all consolidation strengthening.

In figure 3(b), retrieval of all 15 patterns is summed and plotted for different strengths of consolidation. While at lower strengths consolidation is beneficial, at higher strength consolidation becomes detrimental. This is not uniformly the case for all patterns, as figure 3(a) shows. Patterns are therefore grouped, in figure 3(b), into ‘old’ patterns (first five to be stored), ‘middle’ patterns (patterns 6–10) and ‘recent’ patterns (11–15). While ‘old’ patterns benefited from stronger consolidation (especially the first pattern), ‘middle’ and ‘recent’ patterns suffered, leading to lower overall performance with stronger consolidation.

2.4. Discussion

The TraceLink model exemplifies the danger of runaway consolidation that was also shown to occur in a simple Hopfield network. Runaway consolidation is a general problem it will occur whenever the pattern to be consolidated is chosen in a competitive process and when previous consolidation enhances the likelihood that a pattern will win the competition. It is not even limited to connectionist settings. In the first version of a mathematical model of Nadel *et al.*'s (2000) theory of multiple traces, trace replication depended on the number of copies a memory had already assembled. They too were confronted with the phenomenon that a pattern that by chance is replicated often has a higher chance of being replicated later on and in the end swamps the whole memory store (that memory was ‘running away with the replication process’; Nadel *et al.* 2000).

3. Solutions to runaway consolidation

To explore possible solutions to this issue, I took the simulations with the TraceLink model as a starting point. Such solutions should allow the model to function with stronger consolidation (a higher learning parameter), without falling into the trap of runaway consolidation. As the middle patterns in figure 3 (patterns 6–10) benefit from consolidation but not from runaway consolidation, I used average retrieval of these patterns as a measure of successful consolidation.

Runaway consolidation occurs because the likelihood that a pattern is consolidated depends on the amount of consolidation that the pattern has already received. A logical solution would be to attempt to uncouple these two. Another possible solution would be to limit the strength an individual pattern can gain through consolidation, in that way ensuring that the likelihood of consolidation would not vary much between patterns. A third solution would be to try to unlearn too strong attractors, thereby increasing the likelihood that other patterns would be consolidated.

3.1. Suppressed transmission

Runaway consolidation can be seen as a variant of Hasselmo's (1994) ‘runaway synaptic modification’. This is a problem that can plague modellers using unsupervised Hebbian learning. If one lets the weights on connections between two layers determine which connections will encode a new pattern, connections used by one pattern will grow in strength, and therefore be more likely to be used in encoding a second pattern. Ultimately, this tendency will lead to the situation that all patterns are encoded by the same connections. In runaway consolidation too, the weights to be modified determine what attractor is consolidated.

Hasselmo (1999) proposed, as a solution to runaway synaptic modification, a selective suppression of transmission. Transmission in the connections that are modified has to be suppressed, so that their strength does not determine which pattern is stored. Such selective suppression can result from the working of acetylcholine, which has been shown to affect certain inputs more than others in several brain areas.

Selective suppression of transmission in connections to be modified might also be a solution for runaway consolidation. If trace–trace connections were dampened during consolidation, increases in weights resulting from consolidation would not influence the likelihood of a pattern being consolidated in a subsequent trial.

To investigate whether suppression of connections indeed protects the model from runaway consolidation, consolidation strength was varied by setting the learning parameter during consolidation to different values. As in the demonstration of runaway consolidation, total weight change during one consolidation period was set equal to multiples (from 0.5 to 3.5) of the trace acquisition learning parameter. Moreover, suppression of trace–trace connections during consolidation was varied. This was done by multiplying input through trace–trace connections with $1 - \delta$, where δ was set to 0.9 (strong suppression), 0.6, 0.3, or 0 (no suppression).

Replicating results from the previous section (figure 3(b)), retrieval of middle patterns deteriorated with higher consolidation strengths (figure 4, each data point based on 50 replications). With strongly suppressed transmission in trace–trace connections, however, stronger consolidation was beneficial for retrieval. An elimination of runaway consolidation explains this finding. Figure 5 contrasts performance on individual patterns at high consolidation strength (weight change per consolidation period equal to 3.5 times the acquisition learning parameter) for a simulation with no trace–trace connection dampening during consolidation, and strong suppression of trace–trace connections. While in the simulation with no suppression runaway consolidation is evident in the peak performance for the first pattern, no such peak is found in the simulation with strong suppression.

Runaway consolidation is thus eliminated by suppression of transmission in trace–trace connections during consolidation. This occurs because such suppression weakens the connection between the strength of a pattern in trace and the likelihood that it is consolidated.

3.2. Bounded weights

Another way to tackle runaway consolidation is to ensure that patterns do not differ much in the likelihood that they receive a consolidation trial, even after many such trials. One way to do this is to place bounds on trace–trace weights. If the weights subject to

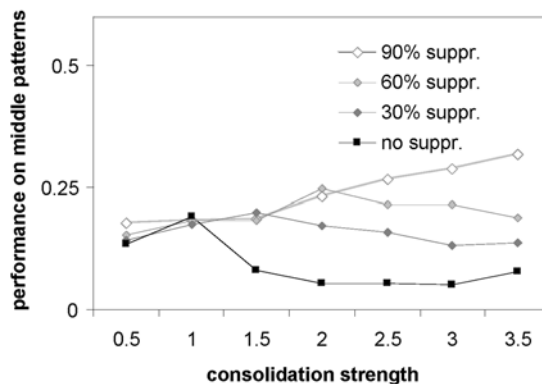


Figure 4. Average retrieval of the middle patterns (patterns 6–10) with different strengths of consolidation and different amounts of transmission suppression in trace–trace connections. Stronger consolidation is detrimental without any suppression of transmission, but beneficial with 60% or 90% suppression of transmission.

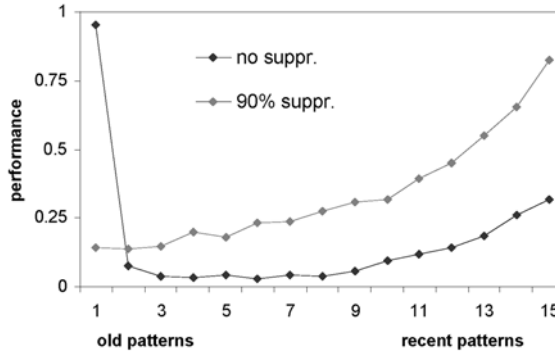


Figure 5. Retrieval probability of the 15 individual patterns with strong consolidation (weight change 3.5 times that of one acquisition trial) and with either maximal suppression of transmission in trace–trace connections (90%), or no suppression of transmission in consolidation. The latter line shows the effects of runaway consolidation.

consolidation cannot grow too large, they cannot strongly bias consolidation toward one pattern. Runaway consolidation is thus made unlikely. This solution was followed by Murre *et al.* (submitted), who rigorously kept weights in all patterns close to their maximal value during the whole simulation.

To investigate this solution, consolidation strength was manipulated as above, but with a bound on the weight of trace–trace connections. A weight was not increased once it reached this bound. Weights were already bounded in all simulations, but to a value of 1 that is typically only reached late in the simulation. Here, the bound on trace–trace connections was set to 1, 0.7, 0.4 or 0.1. The lowest bound of 0.1 is reached after acquisition and two consolidation trials.

As shown in figure 6, lower bounds on trace–trace weights did make the model less sensitive to runaway consolidation; however, it limited strengthening of patterns through consolidation. This is most clear with the smallest bound, 0.1: although patterns did not suffer from stronger consolidation, there were also no benefits. A small increase in performance with higher levels of consolidation was found only with a bound of 0.4.

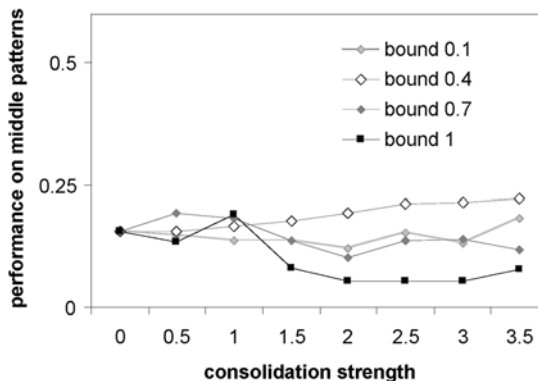


Figure 6. Average retrieval of the middle (patterns 6–10) with different strengths of consolidation and different bounds to trace–trace connections. Higher strengths of consolidation do not harm performance as much with bounds of 0.1 or 0.4 than with higher bounds of 0.7 or 1 (a bound of 1 was in place in all other simulations).

Bounds on the weights subject to consolidation thus eliminate runaway consolidation, but at a price of less effective consolidation.

3.3. Unlearning

Another possible solution to runaway consolidation might be to ‘unlearn’ patterns that gain too much strength in consolidation. Crick and Mitchison (1983, 1995) proposed that unlearning takes place in the brain during REM sleep, in order to remove spurious attractors. This could occur, they thought, next to a consolidation phase taking place during non-REM sleep (Crick and Mitchison 1995). In Hopfield networks, unlearning has been shown to increase the number of patterns that can be retrieved from a network (Hopfield *et al.* 1983, Christos 1996). Although consolidated patterns are usually not spurious patterns but stored ones (Meeter and Murre, submitted), unlearning may still be useful as it weakens too strong attractors, enabling smaller attractors to reemerge. This may also be partly why unlearning works in the Hopfield setting, where strong stored patterns are often the ones that are unlearned (Christos 1996).

To investigate the usefulness of unlearning, a set of simulations was done in which each consolidation trial was followed by an unlearning trial, in which an attractor was found by letting the model iterate for 100 steps, and ‘unlearning’ the attractor unearthed by the model (i.e. learn with a negative learning parameter). As with consolidation, the learning parameter μ in unlearning was set to a multiple of the learning parameter with which a pattern is stored (the trace acquisition learning strength). For example, in one combination tested (labelled as consolidation 1, unlearning -0.5 in figure 7), weights were increased during consolidation with a μ that was one-third of acquisition learning strength, so that if one pattern were consolidated in all three consolidation trials of a consolidation period, its weights would have increased by the same amount as during its initial acquisition. In similar vein, weights were decreased during unlearning with a μ that was one-sixth of acquisition learning strength, so that total unlearning weight change was 50% of acquisition weight change.

Figure 7(a) shows how performance on the middle patterns varies for different consolidation strengths and different levels of unlearning (each data point based on 50 replications). With higher levels of unlearning, there is a slight increase in performance. However, the main effect of unlearning is a shift in peak performance to higher levels of consolidation strength. While without unlearning the maximum performance is at a consolidation strength 1 (equal to acquisition learning strength), the maximum is at 1.5 with an unlearning of -0.5 times acquisition learning strength, at 2 with an unlearning of -1 times acquisition strength, etc.

The reason for these results is simple as the procedure for finding attractors is the same for consolidation and unlearning, the same attractor might first be consolidated and then unlearned (in Christos’s 1996 simulation, just-stored attractors were often targets of unlearning). This would be equivalent to consolidating with a lower learning parameter. Indeed, the difference between learning and unlearning μ s explains most of the variance in figure 7(a). Nevertheless, there is a slight increase in maximal performance with higher levels of unlearning, as is shown in figure 7(b).

A way to increase the beneficial effect of unlearning would be to reduce the likelihood that the same attractors are consolidated and unlearned. This can occur, for example, by suppressing transmission in the connections between trace and link during unlearning. In this way, unlearning is made more dependent on the weights within trace, while both trace and link contribute to the attractor search during consolidation. Indeed, suppressing these connections by 90% led to somewhat better performance with higher levels of unlearning (see figure 7(b)).

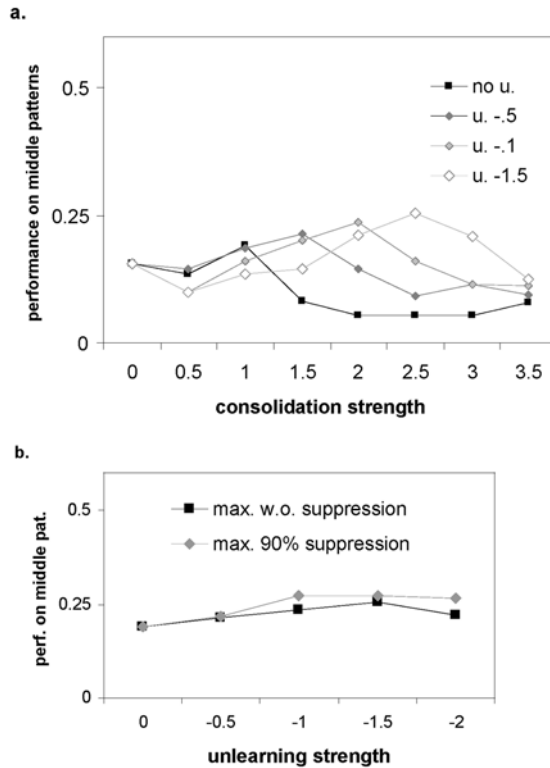


Figure 7. (a) Average retrieval of the middle patterns (patterns 6–10) with different strengths of consolidation and different strengths of unlearning. Strength of consolidation is given in multiples of the trace acquisition learning parameter, strength of unlearning as negative multiples of the same parameter. (b) Maximum retrieval of middle patterns at different strengths of unlearning. Generally, the maximum falls at higher strengths of consolidation for higher strengths of unlearning. Suppressing transmission in trace–link connections improves the efficacy of unlearning, in that performance increases more with higher strengths of unlearning.

4. Discussion

4.1. Consolidation in neural networks

Consolidation has been proposed as a solution to several outstanding problems in connectionism and neuroscience. When implemented in neural networks, it has taken the form of either of two variants of interleaved learning: straight rehearsal (Ratcliff 1990, Murre 1992, McClelland *et al.* 1995, Robins 1995) or pseudorehearsal (Alvarez and Squire 1994, Murre 1996, Robins 1996, Robins and McCallum 1998, 1999, Ans and Rousset 2000, Wittenberg *et al.* 2002, Meeter and Murre, submitted, Murre *et al.*, submitted). Although each implementation has been studied by several researchers, not many of these have implemented in a network how patterns or pseudoitems are stored for consolidation, and how it is decided which pattern or pseudoitem receives a consolidation trial.

When these things are implemented, consolidation may be susceptible to runaway consolidation (Meeter and Murre, submitted). Runaway consolidation is not limited to the TraceLink model, it also appears in a Hopfield network (Wittenberg *et al.* 2002,

Murre *et al.*, submitted) and even appears in a mathematical formulation of consolidation (Nadel *et al.* 2000). The main reason for runaway consolidation is that the likelihood that an attractor emerges from a network increases when that attractor has been consolidated previously. It can be seen as a variant of Hasselmo's (1994) 'runaway synaptic modification', associated with Hebbian learning. The results may not apply to networks in which pseudorehearsal or rehearsal are used in combination with error-correcting forms of learning such as back-propagation, as it is not *a priori* clear that training such a network on one input–output combination increases the likelihood that a similar output is elicited by a different input.

Not all implementations of consolidation have a set-up as shown in figure 1. In some instantiations the training system and the target memory are one and the same. This does not preclude runaway consolidation, as the simulation with the Hopfield network showed. A more elaborate set-up is that of Ans and Rousset (1997, 2000): two coupled networks function as each other's training system, in what they called 'self-refreshing', a form of pseudorehearsal. No runaway consolidation was reported in this network. However, the simulations reported contain only one pass back and forth between the networks. If there were several passes in which one network trains the second and vice versa, old pseudorehearsal trials would have an influence on later pseudorehearsal trials (by influencing the formation of new pseudoitems), and runaway consolidation could be expected to occur as Hebbian learning is used to create input patterns.

Three solutions to runaway consolidation have been discussed in this paper, though it is not implied that other solutions do not exist. The first solution uncoupled the likelihood of consolidation from previous consolidation trials by suppressing transmission in the connections to be modified during consolidation. This led to avoidance of runaway consolidation and to large increases in performance with higher levels of consolidation. The second solution involved placing bounds on the connections subject to consolidation so as to limit the possibility of one pattern garnering enough strength to start runaway consolidation. Indeed, runaway consolidation was avoided, but at a cost: patterns did not benefit very much from higher levels of consolidation. The third method used unlearning to weaken harmfully strong attractors. This increased performance with higher levels of consolidation, though not by much. Making the attractor search in unlearning more dependent on the weights modified in consolidation, however, improved unlearning. All three solutions are thus to some extent effective in avoiding runaway consolidation at higher levels of consolidation.

4.2. Consolidation in the brain

One should be cautious about declaring findings from artificial neural networks valid for natural neural networks, and especially to reason that problems hampering our limited artificial networks must also hamper our brains. Nevertheless, the conclusions reached here seem fairly general. If memories are consolidated in the brain, if that involves rehearsal or pseudorehearsal of memories in a sequential way, and if the likelihood with which a memory is rehearsed increases with previous consolidation, then runaway consolidation is a real danger.

If consolidation occurs in the way described just now, it is possible that the brain has adopted one of the solutions explored here. Not all three solutions are equally biologically plausible. Unlearning during REM sleep, with or without consolidation in slow-wave sleep, is often thought of as not very plausible (Robins and McCallum 1999, Vertes and Eastman 2000, Siegel 2001). It is even doubtful that REM sleep has a role in memory at all: total elimination of REM sleep, as produced by a widely used class of

antidepressants—Mono amine oxidase (MAO) inhibitors—does not notably affect memory function in humans (Siegel 2001). Also, unlearning is not a particularly potent way of protecting memories against interference. It has been claimed that benefits from unlearning can also be attained through a simple weight decay (Robins and McCallum 1999).

The other two mechanisms do not involve such large claims on how the brain works. Suppression by neuromodulators of transmission through specific connections has already been identified in several brain areas (Hasselmo 1999). That this would occur in consolidation is therefore not unimaginable. Indeed, it has been suggested that in slow-wave sleep neuromodulator concentrations are ideal for consolidation of memories from the hippocampus to the neocortex (Hasselmo 1999). Setting bounds on the weights is also consistent with what is known about the brain. Long-term potentiation does not grow unlimitedly. On the contrary, potentiation has been shown to decrease in size when it approaches certain asymptotes (Levy *et al.* 1990). Suppression of transmission in synapses within the neocortex relative to those from the hippocampus, or a maximal potentiation of those synapses not far from their value after initial acquisition, might thus both be viable ways in which the brain could avoid runaway consolidation.

Acknowledgement

The author wishes to thank Jaap Murre for stimulating discussions and helpful comments on previous drafts.

References

- Alvarez, R., and Squire, L. R., 1994, Memory consolidation and the medial temporal lobe: a simple network model. *Proceedings of National Academy of Sciences (USA)*, **91**: 7041–7045.
- Anderson, J. R., and Schooler, L. J., 1991, Reflections of the environment in memory. *Psychological Science*, **2**: 396–408.
- Ans, B., and Rousset, S., 1997, Avoiding catastrophic forgetting by coupling two reverberating neural networks. *Comptes-Rendus de l'Académie des Sciences, Série III*, **320**: 989–997.
- Ans, B., and Rousset, S., 2000, Neural networks with a self-refreshing memory: knowledge transfer in sequential learning tasks without catastrophic interference. *Connection Science*, **12**: 1–19.
- Christos, G. A., 1996, Investigation of the Crick–Mitchison reverse-learning dream sleep hypothesis in a dynamical setting. *Neural Networks*, **9**: 427–434.
- Crick, F., and Mitchison, G., 1983, The function of dream sleep. *Nature*, **304**: 111–114.
- Crick, F., and Mitchison, G., 1995, REM sleep and neural nets. *Behavioural Brain Research*, **69**: 147–155.
- French, R. M., 1999, Catastrophic forgetting in connectionist networks. *Trends in the Cognitive Sciences*, **3**: 128–135.
- Hasselmo, M. E., 1994, Runaway synaptic modification in models of the cortex: implications for Alzheimer's disease. *Neural Networks*, **7** (1): 13–40.
- Hasselmo, M. E., 1999, Neuromodulation acetylcholine and memory consolidation. *Trends in Cognitive Sciences*, **3**: 351–359.
- Hebb, D. O., 1949, *The Organization of Behavior* (New York: Wiley).
- Hopfield, J. J., 1982, Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA*, **79**: 2554–2558.
- Hopfield, J. J., Feinstein, D. I., and Palmer, R. G., 1983, 'Unlearning' has a stabilizing effect in collective memories. *Nature*, **304**: 158–159.
- Lechner, H. A., Squire, L. R., and Byrne, J. H., 1999, 100 years of consolidation—remembering Muller and Pilzecker. *Learning & Memory*, **6**: 77–87.
- Levy, W. B., Colbert, C. M., and Desmond, N. L., 1990, Elemental adaptive processes in neurons and synapses: a statistical/computational perspective. In M. A. Gluck and D. E. Rumelhart (eds) *Neuroscience and Connectionist Theory* (Hillsdale NJ: Lawrence Erlbaum).

- Lopes da Silva, F. H., Witter, M. P., Boeijinga, P. H. and Lohman, A. H., 1990, Anatomic organization and physiology of the limbic cortex. *Physiological Reviews*, **70**: 453–511.
- McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C., 1995, Why there are complementary learning systems in the hippocampus and neocortex insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, **102**: 419–457.
- McCloskey, M., and Cohen, N. J., 1989, Catastrophic interference in connectionist networks: the sequential learning problem. In G. H. Bower (ed.) *The Psychology of Learning and Motivation* (New York: Academic Press), pp. 109–164.
- McGeoch, J. A., 1932, Forgetting and the law of disuse. *Psychological Review*, **39**: 352–370.
- Meeter, M., and Murre, J. M. J., in press, Simulating episodic memory deficits in semantic dementia with the TraceLink model. *Memory*.
- Meeter, M., and Murre, J. M. J., submitted, TraceLink: a connectionist model of consolidation and amnesia.
- Muller, G. E., and Pilzecker, A., 1900, Experimentelle Beiträage zur Lehre vom Gedächtnis. *Zeitschrift fuer Psychologie*, **1**: 1–300.
- Murre, J. M. J., 1992, *Categorization and Learning in Modular Neural Network* (Hillsdale NJ: Lawrence Erlbaum).
- Murre, J. M. J., 1996, TraceLink: a model of amnesia and consolidation of memory. *Hippocampus*, **6**: 675–684.
- Murre, J. M. J., Griffioen, A. R., den Dulk, P., and Robertson, I. H., submitted, Why the brain's memories do not vanish a neural network model of continuous self-repair.
- Nadel, L., Samsonovitch, A., Ryan, L., and Moscovitch, M., 2000, Multiple trace theory of human memory: computational, neuroimaging and neuropsychological results. *Hippocampus*, **10**: 352–368.
- Ratcliff, R., 1990, Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological Review*, **97**: 285–308.
- Ribot, T., 1881, *Les Maladies de la Memoire* (Paris: Germer Baillare).
- Robertson, I. H., and Murre, J. M. J., 1999, Rehabilitation from brain damage: brain plasticity and principles of guided recovery. *Psychological Bulletin*, **125** (5): 544–575.
- Robins, A. V., 1995, Catastrophic forgetting, rehearsal, and pseudorehearsal. *Connection Science*, **7**: 123–146.
- Robins, A. V., 1996, Consolidation in neural networks and in the sleeping brain. *Connection Science*, **8**: 259–275.
- Robins, A. V., 1997, Maintaining stability during new learning in neural networks. *IEEE International Conference on Systems, Man & Cybernetics*, Los Alamos.
- Robins, A. V., and McCallum, S., 1998, Catastrophic forgetting and the pseudorehearsal solution in Hopfield type networks. *Connection Science*, **7**: 121–135.
- Robins, A. V., and McCallum, S., 1999, The consolidation of learning during sleep: comparing the pseudorehearsal and unlearning accounts. *Neural Networks*, **12**: 1191–1206.
- Siegel, J. M., 2001, The REM sleep-memory consolidation hypothesis. *Science*, **294**: 1058–1063.
- Squire, L. R., 1992, Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychological Review*, **99**: 195–231.
- Squire, L. R., and Alvarez, P., 1995, Retrograde amnesia and memory consolidation a neurobiological perspective. *Current Opinion in Neurobiology*, **5**: 169–175.
- Squire, L. R., Cohen, N. J., and Nadel, L., 1984, The medial temporal region and memory consolidation: a new hypothesis. In H. Weingarter and E. Parker (eds) *Memory Consolidation* (Hillsdale NJ: Lawrence Erlbaum), pp. 185–210.
- Treves, A., and Rolls, E. T., 1994, Computational analysis of the role of the hippocampus in memory. *Hippocampus*, **4**: 374–391.
- Vertes, R. P., and Eastman, K. E., 2000, The case against memory consolidation in REM sleep. *Behavioral and Brain Sciences*, **23**: 867–876.
- Wittenberg, G. M., and Tsien, J. Z., 2002, An emerging molecular and cellular framework for memory processing by the hippocampus. *Trends in Neuroscience*, **25**: 501–505.
- Wittenberg, G. M., Sullivan, M. R., and Tsien, J. Z., 2002, Synaptic re-entry reinforcement based network model for long-term memory consolidation. *Hippocampus*, **12**: 637–647.