Chapter 1

PROMOTING BETTER GENERALISATION IN MULTI-LAYER PERCEPTRONS USING A SIMULATED SYNAPTIC DOWNSCALING MECHANISM

A. Brabazon, A. Agapitos and M. O'Neill

Anthony Brabazon
Complex Adaptive Systems Laboratory and
School of Business
University College Dublin
Dublin, Ireland,
anthony.brabazon@ucd.ie

Alexandros Agapitos
Complex Adaptive Systems Laboratory and
School of Computer Science and Informatics
University College Dublin
Dublin, Ireland
alexandros.agapitos@ucd.ie

Michael O'Neill
Complex Adaptive Systems Laboratory and
School of Business
University College Dublin
Dublin, Ireland
m.oneill@ucd.ie

Abstract

A key concern when training a multi-layer perceptron (MLP) is that the final network should generalise well out-of-sample. A considerable literature has emerged which examines various aspects of this issue. In this study we draw inspiration from theories of memory consolidation in order to develop a new methodology for training MLPs in order to promote their generalisation capabilities. The *synaptic homeostasis hypothesis* [32, 33] proposes that a key role of sleep is to downscale synaptic strength to a baseline level that is energetically sustainable. As a consequence, the hypothesis suggests that sleep acts not to actively strengthen selected memories but rather to remove irrelevant memories. In turn, this lessens spurious learning, improves the signal to noise ratio in maintained memories, and therefore produces better generalisation capabilities. In this paper we describe the synaptic homeostasis hypothesis and draw inspiration from it in order to design a 'wake-sleep' training approach for MLPs. The approach is tested on a number of datasets.

1. Introduction

A key concern when applying powerful machine learning methods such as MLPs to induce a model from a training dataset, is that the resulting model should generalise well out of sample. There are several issues that will impact on the generalisation capability of a MLP, including the sufficiency of the training dataset (i.e. does it contain sufficient explanatory inputs in order to allow construction of a predictive model for the target output), is the training data sufficiently representative of all out of sample data that could be presented to the model, is the target function smooth (non-smooth functions will be more difficult to model), and what choice of error criterion will promote good generalisation?

Another factor which will impact on how well an MLP will generalise is its internal structure. If too-large a network is employed, it will have many weights and will be prone to over training, thereby learning any 'noise' in the data. Increasing the number of weights will also add to the computational complexity of the training process. If too-small a network is used, it will not have sufficient power to adequately represent the structure in the data.

Of course, the importance of generalisation extends far beyond machine learning and statistics, and the ability to generalise from past learning to new situations is a key driver of evolutionary fitness in biological organisms. Hence, processes of learning, memory formation, and the integration of new experiences into existing memories in animals, are likely to be rich sources of inspiration for the design of algorithms with good generalisation capabilities.

It is widely thought that iterated wake-sleep states play an important role in memory formation and maintenance in animals. Despite the rich literature in neural networks concerning generalisation, relatively little attention has been paid to the possibility of drawing inspiration from iterated wake-sleep states in order to design better training algorithms for neural networks.

1.1. Memory

Broadly speaking, learning can be considered as the process of acquiring new information, with memory referring to the persistence of learning in a state that can be revealed at a later time [30]. The processes of learning and memory formation have

been widely studied in the literature of both psychology and neurobiology. In the latter case, the focus of research is on how memories are recorded and maintained in the physical structure of the brain. The basic structural unit of the brain consist of individual neurons, a critical aspect of learning and memory is that the connection structure between these neurons is plastic and is altered via the process of learning. The concept of plasticity was first suggested over a century ago by William James [10], and the *synaptic plasticity hypothesis* lies at the centre of most research on memory storage [23]. This hypothesis proposes that the strength of synaptic connections between neurons, which in turn determine the ease with which an action potential in one cell excites or inhibits its target cell, are not fixed but are modifiable or 'plastic'.

While there are multiple types of neurons, the canonical model of information flow at a neuron (the 'neuron doctrine') is that the cell body of a neuron integrates the electrical signals which enter the cell through nerve fibres called dendrites. If the total input signal into a neuron in a time period exceeds a threshold level, the neuron 'fires' and sends an output electrical signal along its axon. In turn, the axon of a neuron is connected to the dendrites of other neurons. Consequently, the firing of a neuron can result in a cascade effect.

A neuron typically has a dense web of input dendrites and these connect, via a synapse, to axon terminals of other neurons at small structures known as dendritic spines. These spines can grow or shrink and are constantly extending out of and retracting back into the dendrite. Hence, the precise network of connections between neurons in a brain is not fixed, but dynamically alters over time. Indeed, two individual neurons may have multiple and not just a single connection. As learning takes place, the network of connections adapts and changes take place at synaptic junctions which can enhance or reduce the ease with which electrical signals can cross the synaptic gap. Memory is stored in a network of linked neurons.

1.1.1. Memory Consolidation

The *memory consolidation hypothesis* was first proposed over a century ago by Müller and Pilzecker [18] and posits that new memories are initially fragile and are only gradually consolidated into long term memory. As noted by [17], while storage of new events in memory can occur very quickly (within seconds), slow consolidation of memories into long term storage (a process which can take days, weeks, or even longer) may be adaptive as it allows for a dynamic interplay between current experience and pre-existing memories.

The term memory consolidation is itself variously defined as, 'a time-dependent, off-line process that stabilizes memories against interference and decay, allowing them to persist over time' [17], a 'process that transforms new and initially labile memories encoded in the awake state into more stable representations that become integrated into the network of pre-existing long-term memories' [5], or as 'the processes that stabilise the learning-induced changes in synaptic morphology that represent the biological substrate of memory' [8].

In discussing memory consolidation, a distinction is drawn between:

1. cellular consolidation, and
2. systems consolidation.

Cellular consolidation arises from a series of biochemical events which take place in individual synapses, typically within a short time frame (minutes to hours) after the initial experience. System consolidation refers to events which take place over a

longer time frame and which are thought to maintain the memory in long term memory storage.

Rudy (2014) [23] provides an excellent review of the current state of understanding of how memories are created and maintained. While there is still considerable debate concerning several aspects of this process, the most widely accepted view is that memory develops over a number of stages namely, generation, stabilisation, consolidation and maintenance.

Initially, there are changes in the synaptic strength of the effected neurons, resulting from a reorganisation of existing proteins in the relevant dendritic spine and axon terminal. For example, within minutes, the number of glutamate receptors in the spine is increased thereby facilitating the enhanced transmission of sodium ions (electrical signal) between the axon terminal and the spine. To consolidate the synaptic change further, in following hours transcription and translation processes are activated creating new proteins. These have several effects including the enhancement of the degree of bonding between the spine and axon, and an alteration of the physical geometry of the spine. This further promotes the transmission of ions between the spine and axon. Typically, this process lasts for up to 24 hours and helps ensure that the physical changes in the synapses endure for several days.

While the above explains how synaptic changes initially occur and are subsequently stabilised, it does not explain how strengthened synapses that support memory outlive the molecules from which they are made. This is known as the 'molecular turnover problem' and is a active area of research inquiry. In order to maintain a memory, a variety of proteins need to be continually manufactured at the synapse, even in the absence of the original stimulus. Recent work by **??** suggests that self-sustaining (self-copying) populations of proteins may be the key to maintaining the long-term synaptic changes that underlie memory.

Obviously, there is little reason to maintain a memory of most of the routine events which occur during a day, and indeed experience suggests that we will forget much of this detail within several days. It is speculated that memories are most likely to be maintained for the long term when either the behavioural experience is considered significant, is repeated, or when the memory is recalled [8]. As will be discussed later, it is thought that sleep plays an important role in long term memory consolidation.

1.1.2. Memory Systems

When discussing memory, is important to note that the the brain has multiple memory systems, depending on the nature of what is being learnt. Perhaps the best known system is that for declarative memory which includes both episodic memory (memory for facts and events) and semantic memory (supports memory for facts and provides an ability to generalise from multiple experiences). This system relies on an interplay between the neocortex, the hippocampus and its related cortical structures. Sensory information passes into the neocortex and in turn is processed and passed via a number of intermediate structures into the hippocampus. By the time the information passes into the hippocampus it is already highly processed and amodal (hippocampus neurons do not know whether they are receiving auditory, visual or other sensory inputs) [23].

Although it is known that the hippocampus plays a vital role in episodic memory, there is debate as to how exactly it does this. One theory is the 'indexing theory of episodic memory' [35]. According to this theory, the content of episodic memories are stored in the neocortex and the hippocampus creates indices to these memories by binding the inputs it receives from the different regions of the neocortex into a neural ensemble that represents the conjunction of their co-occurance [23]. The hippocampus

projects back to the neocortex when the index is activated.

In essence, the theory assumes that events create a memory trace by activating patterns of neocortical activity, which then project to the hippocampus, with the relevant synapses in the hippocampus responding to the neocortical inputs being strengthened via long term potentiation (LTP). Therefore, the hippocampus acts as an index to a 'memory' filing cabinet which enables the recall of memories, even when only a subset of the original neocortical pattern is received by the hippocampus. Although this may appear to be an unneccessarily complex process, it is posited that it may have arisen due to structural limitations of the neocortex as potential associative connectivity across neocortical regions is low [23]. It is also speculated that memories in the neocortex may potentially have more than one index associated with them, if the event is repeated or if the memory is reactivated (recalled). Hence, the more often an item is experienced or recalled, the more 'paths' to it may be generated in the hippocampus. This is known as the *multiple trace theory* [19].

1.2. Sleep and Memory Consolidation

At first glance being asleep would appear to be a potentially dangerous and costly activity as sleeping animals cannot forage for resources, take care of young, procreate, and are exposed to predation risk [7, 14]. Despite these drawbacks, sleep behaviours are widespread in the animal kingdom and it is evident that many animals spend a significant portion of their day in sleep or in sleep-like states. Evolution has even devised some extraordinary adaptations to accommodate sleep [34]. Perhaps the most unusual of these adaptations is exhibited by cetaceans (including whales, dolphins and porpoises) who can engage in unihemispherical (or 'half-brain') sleep, wherein one eye is kept open during sleep, with the contralateral side of the brain also remaining awake [21]. Other examples of unihemispheric sleep include some species of birds [22] which can keep one eye open during sleep, particularly if the predation risk is high.

Given the widespread nature of sleep behaviour, and the lengths to which evolution has gone in order to conserve sleep in some animals, one could well ask what benefit does sleep provide that makes it crucial to living creatures?

Amongst the multiple potential functions of sleep, one of the most heavily researched is whether sleep plays a role memory formation and maintenance. In many species, the same regions of the brain that process sensory information are also important for memory formation. This poses a dilemma, as if these regions are busy processing sensory information during waking, then it is likely to be more difficult for processes such as memory consolidation to take place simultaneously, in turn leading to a suggestion that sleep may allow these conflicting activities to co-exist, leading to a claim that memory consolidation occurs predominately during sleep [1]. .

In this study we draw inspiration from the synaptic homeostasis hypothesis which is drawn from the literature on memory consolidation in order to design a training approach for an MLP which is capable of generalising from noisy data. Therefore, we simulate a wake-sleep cycle during which the MLP is presented with new sensory inputs (data) during the wake phase, leading to synaptic potentiation, with synaptic downscaling taking place during a simulated 'sleep' phase. Critically and in contrast to prior literature on weight-decay processes for training of MLPs, the training process takes place over a sequence of simulated wake-sleep phases.

1.3. Structure of Paper

The remainder of this paper is organised as follows. Section 2 provides some background on two theories of memory consolidation during sleep. Section 3 describes the model developed in this paper and outlines the experiments undertaken. The results of these are presented and analysed in section 4, with conclusions and suggestions for future work being presented in section 5.

2. Background

In this section we provide some background on memory consolidation during sleep, and in particular, we describe the synaptic homeostasis hypothesis. We also overview some previous literature which has applied ideas from the process of memory consolidation for neural network training.

2.1. Sleep States

A common way to characterise sleep state is to examine the electrical activity of the brain recorded using an electroencephalogram (EEG). In mammals and birds sleep can be divided into two main phases namely, REM (rapid eye movement) and NREM (non rapid eye movement) sleep. REM sleep is characterised by high frequency, low amplitude, electrical activity in the brain, and this bears some similarity to the electrical activity of the brain during wakefulness. In contrast, NREM sleep is characterised by the propagation of low frequency (slow), high amplitude, electrical waves in the brain.

In humans, NREM sleep is divided into three successive stages [24], and the sleep cycle follows a typical ordering of stage 1 NREM, stage 2 NREM, stage 3 NREM, and finally REM sleep. The entire cycle lasts some 90-100 minutes and repeats itself several times during the night. As the sleep cycles progress, the portion of time spent in NREM sleep reduces and the portion of time in each cycle spent in REM sleep increases. Sleep during stage 3 of NREM sleep is termed slow wave sleep (SWS), and is characterised by delta wave activity brain activity, which produces the lowest frequency and highest amplitude patterns of electrical activity.

2.2. Active System Consolidation Hypothesis

There are currently two hypotheses concerning the mechanisms underlying the consolidation of memory during sleep. The active system consolidation hypothesis (ASCH) proposes that an active consolidation process results from the re-activation of selected memories during sleep [5], and the synaptic homeostasis hypothesis (SHH) assumes that consolidation may also occur during waking and that the role of sleep is to restore the encoding capabilities of synaptic connections (global synaptic downscaling) [1].

The ASCH arose from the standard model of systems consolidation for declarative memory [16]. Different regions of brain are responsible for different memories, with *declarative memory* (these memories are accessible to conscious recollection and include memories for facts and events) relying on the hippocampus and neocortical regions of the brain, and *procedural memory* (memories for skills that result from repeated practice e.g. riding a bike or playing a piano) relying on the striatum and cerebellum [5]. The standard two-stage theory for declarative memory consolidation proposes that there are two separate memory stores. One allows learning at a fast rate and serves as an intermediate buffer to hold information temporarily. The other store

learns at a slower rate and serves as long-term memory. For declarative memory, sensory information in the waking brain flows into the cortex and it is proposed that events are initially encoded in parallel in neocortical networks and also in transient neuronal assemblies in the hippocampus.

Although the theory did not initially outline a role for explicit recall in the consolidation of the long term memory, it has been suggested that during sleep, a two-way dialogue between the hippocampus and neocortex takes place in order to effect memory consolidation [5]. The hippocampus can be considered as a rapidly-encoded, sparse, memory system which allows for the formation of event memories, whereas the neocortex is a slowly-consolidating, dense, memory storage system. During NREM sleep, slow (electrical wave) oscillations, spindles, and ripples coordinate the reactivation and redistribution of hippocampus-dependent memories to neocortical sites. The newly-acquired memory traces are reactivated and it is claimed that information flows from the hippocampus to the cortex, such that connections in the neocortex are strengthened, forming more persistent memory representations. In REM sleep, it is proposed that the information flow reverses (from the neocortex back to the hippocampus). This two-way process iterates during the period of sleep [31], thereby modifying the representations in both stores, and integrating the new memory into pre-existing memories. This enables the extraction of invariant features, including the forming of new associations, and eventually insights into hidden rules and patterns [5]. Hence, through the repeated re-activation of the new memories during sleep, the fast learning store acts as an internal trainer of the slow learning store to gradually adapt the new memories to the pre-existing network of long term memories [5].

There is some evidence to support the ASCH, as we know from brain imaging studies that the spatio-temporal patterns of neuronal firing that occur in the hippocampus, during the exploration of a novel environment or during simple spatial tasks, are reactivated in the same order during subsequent sleep. However, we do not have a detailed understanding as to how these reactivations could stimulate the strengthening of links between neocortical storage sites, and specifically, how enduring synaptic changes could result in the neocortex [5]. In the standard two-stage theory, the consolidation process that takes place off-line relies on the re-activation of the neuronal circuits that were implicated in the initial encoding of the memory, and therefore consolidation involves the *reinforcement* of memory representations at the synaptic level. Long-term potentiation (LTP) (Hebbian learning - the assumption that information is stored in the brain as changes in synaptic efficiency which occur when neurons fire synchronously together) is considered a key mechanism of synaptic consolidation. It is not certain whether memory re-activation during sleep promotes the redistribution of memories by inducing new LTP (at long-term storage sites) or whether re-activation merely enhances the maintenance of LTP that was induced during encoding. An assumption of the traditional two stage model is that LTP takes place in the long term memory store as a result of selective reactivation of memories during system consolidation.

Although we await further investigation of sleep dependent learning, recent work by [39] has indicated that sleep (specifically, NREM sleep) by mice after a motor learning task promoted new spine formation in the motor cortex of those mice.

It has been speculated that spindle oscillations which are concentrated in stage 2 NREM, open molecular gates to plasticity by evoking calcium entry in neocortical pyramidal neurons, priming the neurons for biochemical events that could lead to permanent changes in the network. Consolidation could then proceed by iteratively recalling and storing information in primed neural assemblies [25]. One interesting feature of reactivations during SWS is that they appear to be noisier, less accurate, and often happen at a faster firing rate than the related activity during the initial encod-

ing phases. Plausibly this 'noisy' teaching could result in more robust memory in an analogue to using 'jitter' in training MLPs.

2.3. Synaptic Homeostasis Hypothesis

An alternative perspective which has gained a significant following in recent years is the *synaptic homeostasis hypothesis* (SHH) [32, 33, 34]. This hypothesis suggests that the primary memory function of sleep is to produce a global synaptic downscaling, and that memory consolidation is continuous (i.e. can occur during waking) and not limited to sleeping states.

The proponents of the SHH do not disagree that memories form as neurons that get activated together strengthen their links through synaptic potentiation, nor that brains replay newly-learnt material at night, or that patterns of neural activity during sleep sometimes resemble those recorded while a subject is awake. However they question conventional wisdom that brain activity during sleep reinforces the synapses involved in storing newly-formed memories, noting that there is no strong evidence that synapses in replayed circuits get strengthened during sleep [34]. Instead they claim that a critical driver of sleep is a need to restore the brain to a baseline state, by *weakening* the links between neurons during sleep, in order to preserve the brain's ability to learn and form new memories while it is awake. The weakening process is termed *synaptic downscaling*.

Brain tissue is metabolically expensive. In humans, the brain while accounting for only about 2% of total body mass, consumes some 20% of energy requirements during quiet waking [27]. Approximately 2/3 of this energy consumption goes to supporting and maintaining synaptic activity. Strong synapses consume more energy than weak ones and the energy budget available to brain tissue is not unlimited. During the day, the potentiation of synaptic circuits from sensory inputs results in an increase in the number and size of synapses, leading to a higher level of energy requirement [34]. Advocates of the SHH claim that a generalised depression of synapses during sleep would benefit the brain as it would decrease the energy cost of synaptic activity, eliminate weak and ineffective synapses, and reduce cellular stress [4].

An important part of effective learning is a corresponding 'forgetting' of irrelevant memories. Under the SHH, synaptic potentiation stemming from daytime learning is down regulated brain-wide during slow wave sleep. Crucially, it is assumed that this rescaling process preserves relative synaptic weight differences, and therefore may lead to forgetting because the downscaling may effectively silence, or even remove, spines with synapses that are only weakly potentiated. Down selection under the hypothesis promotes survival of only the fittest neural circuits, either because they were activated strongly and consistently during wakefulness, or because they were better integrated with pre-existing memories (for example, a new word in a known language). Synapses that were only mildly enhanced during wakefulness, or which fit less well with existing memories would be depressed, and leave no lasting trace in our neural circuitry.

While there is experimental evidence for several aspects of synaptic downscaling [22], including evidence from animal studies that the number and size of spines and related synapses reduces during sleep [34], there is as yet no direct evidence for a specific mechanism which selectively weakens activated synapses during sleep [34]. It is speculated that the slow waves of mammalian NREM sleep play a role. We know that at sleep onset, levels of SWA are elevated as a result of synaptic strength accrued during learning while awake. This increase in effective connectivity causes the slow-oscillations of neurons to be more synchronous, and thereby levels of SWA to be

high [22]. The large-scale slow oscillations of neuronal networks may produce synaptic downscaling, a global decrease in synaptic strength, and an increase the signal to noise ratios for important memories by eliminating synapses below a certain threshold. This may explain why performance on certain cognitive tasks increases following sleep [22]. Interestingly, synaptic downscaling is a self-limiting process because as synapses weaken, neurons oscillate less synchronously and consequently induce less downscaling [22](p. 265). It is also known that the chemistry of the brain changes during sleep and Tononi and Cirelli [34] have speculated that this could bias neural circuitry so that synapses becomes weakened rather than strengthened when signals flow across them.

The SHH, with its emphasis on an 'active decay' (forgetting) of irrelevant memories during sleep, provides an interesting alternative to the traditional idea of sleep-mediated synaptic strengthening of important memories. Most memories formed during the day are irrelevant and a decay process which ensured that unwanted and unneeded memories are removed could result in a lessening of spurious learning and better generalisation capabilities [8].

In this study, we do not claim that the SHH provides a more correct description of memory consolidation during sleep than the ASCH as current empirical evidence does not conclusively support the SHH. Indeed, it has been noted by Axmacher et al. [1] and by Diekelmann and Born (2010) [5] that the ASCH and the SHH are not necessarily mutually exclusive, as a sequential process could exist with active system consolidation integrating newly encoded memories with pre-existing long term memories thereby inducing conformational changes in the neocortex followed by global synaptic downscaling in order to avoid the saturation of synaptic networks. Rather, we draw inspiration from the SHH in order to design a training process for MLPs.

2.4. Synaptic Downscaling and Regularisation

The synaptic downscaling concept bears interesting comparison with some classical approaches to regularisation in the neural network literature. Broadly speaking, regularisation is any modification to a learning algorithm which aims to reduce the chance of overfit. Typically, the object is to smooth the response of the final model. Common methods for regularisation include early stopping, wherein training is stopped when the error measure on a hold-out validation sample begins to increase, or the inclusion of a penalty term in the error function for model complexity. In applications of the latter in MLPs, the error metric is usually defined as MSE plus an additional (weighted) term which consists of the sum of the squares of the weights. This alteration to the error function will tend to reduce weight sizes in the final network and therefore make the network's response smoother. In turn this will tend to reduce overfit as over-fitted mappings require high curvature and hence large weights. The general form of the regularised cost function in this case is given by:

$$E_{reg} = E_{mse} + \alpha\omega \tag{1}$$

where α is the regularisation parameter which controls the trade-off between reducing the error and increasing the smoothing. The term ω is a penalty function which captures the complexity of the underlying network. If the penalty is defined as the sum of the squares of the weights in the MLP, the approach bears similarity to ridge regression in linear models, and it effectively implements a form of 'weight decay' as in each epoch individual weights decay in proportion to their previous size, i.e. exponentially, unless the weight is changed in the learning process [20]. A wide number of variants

on this basic approach have been examined including,'weight elimination' [37], where the decay process is tuned in order to shrink small weight coefficients more heavily.

Although even basic weight decay approaches can notably improve generalisation capabilities [12], we cannot assume that is is optimal to apply the same decay constant to all weights in the network, and in particular, we could suppose that different decay constants should be applied to connections between input and hidden, hidden to hidden, and hidden to output nodes. Nor can we assume that it is optimal to apply the same decay constant(s) for the entire training process, and [37] illustrate an approach where the decay constant is iteratively updated during training.

Apart from reducing the values for weight parameters in a network, another way to attempt to improve generalisation is to directly restrict or seek to reduce the structural complexity of the network. This can be done by restricting the number of hidden layer nodes, or by 'pruning' individual node connections in a network. One approach is to set connections with small weights to zero, thereby 'tuning off' or 'pruning' that connection. After the relevant weights are deleted, the (reduced) network is retrained. A significant number of studies applying network pruning have resulted over the past 25 years following early work by [28, 29].

As noted by [13], there is a close link between weight decay and pruning, as an iterated pruning process effectively reduces to continuous weight-decay during training. A downside of these approaches is that the learning process can be slow due to the need for repeated re-training and there is an implicit assumption that deletion of connections with small weight values will not have much effect on model fit. A better, if often computationally prohibitive, approach would be to delete weights, whose deletion will have least effect on training error (or to train the network using all possible subsets of weights [11]). Of course, to determine which weight to delete, the MLP would need to be iteratively retrained with each weight being removed in turn.

A more computationally feasible approach to pruning was developed by Lecun et al. (1989) [13], namely the optimal brain damage (OBD) approach. In OBD the second derivatives of each weight parameter with respect to the error function are used in order to determine which weights to remove. As for other pruning methods, OBD proceeds in an iterative manner. Initially the full network is trained on the data, a pruning process is then applied, and the new network is then retrained.

From the above discussion, we can see that weight decay and pruning, both features of the SHH, are well-developed techniques in the neural network literature. It is interesting to note that the development of these techniques stemmed from a statistical rather than a biological perspective. An important aspect of the memory consolidation process that has not yet been embedded in the regularisation literature is the iterative nature of memory consolidation, with new memories only being slowly integrated into existing knowledge, with both memories being altered in this process.

2.5. Neural Network Derived from a Sleep Metaphor

As noted in section one, relatively little attention has been paid to the use of 'sleep' metaphors for design of neural network algorithms. Perhaps the best known of these algorithms is the 'wake-sleep' algorithm of Hinton et al. [9] for unsupervised learning which draws on the standard model of systems consolidation for declarative memory. In Hinton's study, a multilayer network of simulated stochastic neurons is described, with bottom up recognition connections during the wake phase being used to produce a representation of inputs in one or more hidden layers. In the 'wake' phase, neurons are driven by recognition connections, and generative connections are adapted to increase the probability that they would reconstruct the correct activity vector in

the layer below. In the 'sleep' phase, neurons are driven by generative connections, and recognition connections are adapted to increase the probability that they would produce the correct activity vector in the layer above. By alternating activity in two directions, the hidden layer representations are modified until they produce an optimal representation of the original signal.

3. Model and Experiments

The general model we adopt for our experiments is a feed forward multi-layer perceptron (MLP). We create training data from four test functions, and for each input vector in the training set, we inject differing amounts of noise into the associated function output, thereby producing 'learning' problems of varying difficulty.

The MLP is exposed to a succession of non-overlapping 'windows' of training data during its wake cycles. During exposure to each training vector, a learning process takes place in which synaptic potentiation via the back propagation training algorithm is simulated. At the end of each data window, a sleep cycle is simulated during which synaptic downscaling takes place, and this in turn is followed by another wake cycle in which a new window of training data is presented to the MLP network. During downscaling each weight is decreased by a certain percentage.

Once the MLP has been trained, its out of sample performance on clean test data, generated using the relevant function, is assessed. This allows us to determine how well the MLP has performed in uncovering the correct underlying function, in spite of being presented with noisy data during training.

The results from the MLP developed using a simulated synaptic downscaling process are benchmarked against those produced by a feed forward MLP which has been trained in one pass over the training data.

3.1. Datasets

We selected a suite of four synthetic regression problems so that we can reliably generate data with specific amounts of noise. Figures 6, 7, 8 are graphical representations of the bivariate problems F_1m F_3, and F_4 respectively. In every synthetic dataset, we randomly sample 100 training examples of the form (x, y), where the input vector $x \in R^d$, and the response variable $y \in R$. The goal is to learn a target function f that maps x to y. The response variable of each example is corrupted by random noise drawn according to a Gaussian probability distribution with certain μ and σ. Thus each training set of examples takes the form $\{(x_i, z_i)\}_1^{100}$, where $z_1 = f(x_i) + e_i$. $f(x_i)$ is the noise-free value of the target function and e_i is a random variable representing the noise. We experiment with a set of σ values defined as $\{0.01, 0.1, 1.0, 10.0, 30.0, 50.0\}$, and μ set to 0.0. The details of the sampling procedure used for generation of training and test data for different problems are given in Table 1. Note that noise is only added to the training data, whereas the data used to assess model generalisation is not contaminated.

Furthermore, the response value in each input-output pair is normalised within the [0.0, 1.0] interval prior to training. Normalisation of a noise-corrupted value α is performed using $(\alpha - min)/(max - min)$, where min and max are the minimum and maximum values out of 100 training response values respectively. Figure 5(a) shows the histograms of the normalised response values for different regression problems. The same normalisation applies to testing data, however this time each response value

Table 1. Regression problems with the respective data sampling ranges for training and test datasets. Notation x=rand(a,b) means that the x variable is sampled uniform randomly from the interval [a, b].

	Problem	Training data	Test data
F_1	$f(x_1, x_2) = \frac{e^{-(x_1-1)^2}}{1.2+(x_2-2.5)^2}$	100 points x_1, x_2=rand(-3.0, 3.0)	10,000 points x_1, x_2=rand(-3.0, 3.0)
F_2	$f(x_1, x_2, x_3, x_4, x_5) = \frac{10}{5+\sum_{i=1}^{5}(x_i-3)^2}$	100 points x_1, x_2, x_3, x_4, x_5 =rand(-3.0, 3.0)	10,000 points x_1, x_2, x_3, x_4, x_5 =rand(-3.0, 3.0)
F_3	$f(x_1, x_2) = x_1 * x_2 + sin((x_1 - 1) * (x_2 - 1))$	100 points x_1, x_2=rand(-3.0, 3.0)	10,000 points x_1, x_2=rand(-3.0, 3.0)
F_4	$f(x_1, x_2) = \frac{(x_1-3)^4+(x_2-3)^3-(x_2-3)}{(x_2-2)^4+10}$	100 points x_1, x_2=rand(-3.0, 3.0)	10,000 points x_1, x_2=rand(-3.0, 3.0)

is noise-free. Figure 5(b) shows the histogram of the normalised response values for different problems.

3.2. MLP design

The regression problems F_1, F_2, F_3, F_4 are of two, five, two and two input variables respectively. The architecture of a MLP consists of an input layer with the same number of input nodes as the dimensionality of the input of a problem, a hidden layer of 10 nodes with *tanh* activation functions, and an output layer of a single node with a *tanh* activation function. Training is performed using standard back-propagation with a learning rate set to 0.005, iterated for $2,000$ epochs. We are experimenting with the effect of the number of wake-sleep cycles during training, and tried the proposed method with 5, 10, 15, and 20 cycles. This effectively means that each set of training examples is divided into the respective number of non-overlapping subsets.

4. Results and Analysis

In the figures discussed in this section, we plot the average Mean Squared Error (MSE) that accrues from 50 runs of an MLP using different random weight initialisations as a function of the weight downscaling percentage that takes place during a sleep phase. For comparison purposes we also plot the average MSE that is obtained from the baseline MLP algorithm that uses no weight downscaling. Depending on the level of noise that is injected into each response variable, we categorise the learning problems into easy (Gaussian noise σ of 0.01 or 0.1), moderate (σ of 1.0 or 10.0), and hard (σ of 30.0 or 50.0). In addition, Tables 2, 3, 4, 5 present the standard errors for the out-of-sample MSE estimates.

Figure 1 presents the results for problem F_1. An observation that is consistent across all different setups for the number of wake/sleep cycles is that for the easy and moderate problem formulations the proposed method outperformed standard MLP. In addition, results suggest no clear trend in the evolution of the MSE curve as a function of the downscaling percentage, however for the smaller levels of noise (i.e. 0.01, 0.1) increasing the percentage of downscaling seems to worsen the generalisation

performance.

The results for problem F_2 are presented in Figure 2. Here the number of wake/sleep cycles exert an effect in the out-of-sample performance with their number set to 15 attaining the best generalisation improvement over standard MLP for all problem formulations but the one where noise σ is set to 1.0. Results also suggest that in the easiest case (i.e. noise σ of 0.01), the method of downscaling is difficult to improve performance over standard MLP and in most cases leads to performance deterioration.

Figure 3 presents the results for problem F_3. In this case, contrary to the results observed in other problems, weight downscaling improves performance over standard MLP in the least noisy problems, whereas the performance deteriorates over that of standard MLP for the noisiest problem formulation. This increase in performance in the case of noise levels of 0.01 and 0.1 can be attributed to the discrepancy between the distributions of the response values between training and testing as can been seen in Figure 5(a) for Function 3 under noise level 0.01 and Figure 5(b) for the same function. This was due to random sampling for the values x_1 and x_2 that created relatively disjoint sets of examples to train and test a model. The median of the response values is approx. 0.03 for training, and 0.57 for testing. Out-of-sample performance is therefore improved by relaxing the fit to the training examples. This incidental result should be regarded as a valid scenario of training and testing data distribution mismatch that can occur when dealing with real-world data. It reinforces the view that in case of overfit models, weight downscaling can improve out-of-sample performance.

Finally, Figure 4 presents the results for problem F_4. We observe that the use of downscaling substantially improves the out-of-sample performance for the noisiest problem formulations. This is evident in the case were the number of wake/sleep cycles was the greatest, i.e., 15 and 20. For the easy and moderate cases, figures suggest that downscaling has the tendency to worsen performance. This particular problem also exhibits an interesting trend in the evolution of the the MSE curve as a function of the downscaling percentage. More specifically, the out-of-sample error decreases as the downscaling percentage increases for the noisiest problems, whereas it decreases as a function of increasing percentage of the small and moderate levels of noise in the target.

4.1. Summary of observations

The observations from the experiment can be summarised as follows:

1. The downscaling mechanism increases the generalisation performance for most cases of moderate and high levels of noise.
2. No advantage is accruing from the proposed method when used with small levels of noise in the target function. In most cases, performance deteriorates.
3. The optimal number of wake/sleep cycles and the level of weight downscaling appears to be problem dependent. A principled approach such as cross-validation should be applied to chose these effectively.
4. Overall, when training and testing over similar input-output distributions, weight downscaling exerts a negative effect by disrupting the fit of a model. On the other hand, in the case where there is discrepancy between training and testing input-output distributions, the downscaling mechanism improves generalisation.

Table 2. Out of sample MSE: mean values and standard errors. Function 1.

Down	Noise 0.01	Noise 0.1	Noise 1.0	Noise 10.0	Noise 30.0	Noise 50.0
			Standard MLP			
n/a	0.060 (0.0005)	0.077 (0.0011)	0.151 (0.0005)	0.195 (0.0012)	0.164 (0.0013)	0.172 (0.0009)
			5 wake/sleep cycles			
1%	0.063 (0.0007)	0.068 (0.0016)	0.166 (0.0011)	0.194 (0.0028)	0.245 (0.0023)	0.169 (0.0036)
15%	0.062 (0.0005)	0.069 (0.0008)	0.162 (0.0013)	0.158 (0.0018)	0.252 (0.0022)	0.173 (0.0018)
30%	0.067 (0.0003)	0.074 (0.0009)	0.159 (0.0016)	0.158 (0.0009)	0.239 (0.0030)	0.160 (0.0010)
			10 wake/sleep cycles			
1%	0.061 (0.0027)	0.077 (0.0018)	0.123 (0.0029)	0.128 (0.0029)	0.218 (0.0036)	0.160 (0.0049)
15%	0.084 (0.0038)	0.061 (0.0003)	0.127 (0.0023)	0.122 (0.0008)	0.201 (0.0020)	0.186 (0.0030)
30%	0.105 (0.0015)	0.069 (0.0003)	0.134 (0.0018)	0.135 (0.0012)	0.214 (0.0019)	0.181 (0.0017)
			15 wake/sleep cycles			
1%	0.056 (0.0021)	0.064 (0.0014)	0.159 (0.0035)	0.121 (0.0032)	0.228 (0.0019)	0.154 (0.0045)
15%	0.078 (0.0032)	0.066 (0.0009)	0.139 (0.0005)	0.111 (0.0009)	0.232 (0.0019)	0.215 (0.0004)
30%	0.099 (0.0004)	0.077 (0.0008)	0.141 (0.0004)	0.111 (0.0004)	0.206 (0.0009)	0.198 (0.0003)
			20 wake/sleep cycles			
1%	0.052 (0.0015)	0.078 (0.0008)	0.074 (0.0017)	0.182 (0.0044)	0.239 (0.0077)	0.167 (0.0022)
15%	0.074 (0.0004)	0.077 (0.0008)	0.072 (0.0004)	0.195 (0.0016)	0.243 (0.0027)	0.212 (0.0005)
30%	0.086 (0.0004)	0.101 (0.0005)	0.081 (0.0002)	0.188 (0.0005)	0.234 (0.0009)	0.206 (0.0002)

Table 3. Out of sample MSE: mean values and standard errors. Function 2.

Down	Noise 0.01	Noise 0.1	Noise 1.0	Noise 10.0	Noise 30.0	Noise 50.0
			Standard MLP			
n/a	0.034 (0.0007)	0.107 (0.0019)	0.167 (0.0027)	0.174 (0.0057)	0.204 (0.0073)	0.234 (0.0041)
			5 wake/sleep cycles			
1%	0.188 (0.0053)	0.141 (0.0041)	0.203 (0.0127)	0.202 (0.0117)	0.156 (0.0042)	0.212 (0.0126)
15%	0.087 (0.0040)	0.074 (0.0013)	0.186 (0.0072)	0.169 (0.0074)	0.149 (0.0026)	0.225 (0.0077)
30%	0.031 (0.0009)	0.064 (0.0022)	0.197 (0.0040)	0.149 (0.0036)	0.147 (0.0019)	0.226 (0.0058)
			10 wake/sleep cycles			
1%	0.124 (0.0086)	0.103 (0.0043)	0.263 (0.0090)	0.375 (0.0141)	0.082 (0.0047)	0.289 (0.0115)
15%	0.025 (0.0008)	0.077 (0.0017)	0.234 (0.0030)	0.238 (0.0037)	0.068 (0.0027)	0.206 (0.0060)
30%	0.035 (0.0007)	0.083 (0.0005)	0.255 (0.0016)	0.169 (0.0014)	0.065 (0.0015)	0.223 (0.0017)
			15 wake/sleep cycles			
1%	0.048 (0.0033)	0.107 (0.0053)	0.296 (0.0050)	0.175 (0.0087)	0.084 (0.0021)	0.202 (0.0095)
15%	0.038 (0.0012)	0.075 (0.0007)	0.269 (0.0004)	0.066 (0.0012)	0.067 (0.0026)	0.172 (0.0031)
30%	0.043 (0.0007)	0.087 (0.0006)	0.266 (0.0005)	0.086 (0.0005)	0.061 (0.0028)	0.218 (0.0008)
			20 wake/sleep cycles			
1%	0.037 (0.0009)	0.083 (0.0026)	0.124 (0.0054)	0.149 (0.0062)	0.307 (0.0107)	0.286 (0.0064)
15%	0.042 (0.0004)	0.102 (0.0007)	0.150 (0.0007)	0.108 (0.0005)	0.236 (0.0024)	0.249 (0.0006)
30%	0.054 (0.0002)	0.117 (0.0009)	0.187 (0.0001)	0.122 (0.0003)	0.178 (0.0009)	0.229 (0.0005)

Table 4. Out of sample MSE: mean values and standard errors. Function 3.

Down	Noise 0.01	Noise 0.1	Noise 1.0	Noise 10.0	Noise 30.0	Noise 50.0
			Standard MLP			
n/a	0.203 (0.0002)	0.095 (0.0003)	0.030 (0.0001)	0.037 (0.0001)	0.050 (0.0007)	0.036 (0.0001)
			5 wake/sleep cycles			
1%	0.203 (0.0010)	0.113 (0.0037)	0.043 (0.0007)	0.043 (0.0005)	0.077 (0.0019)	0.053 (0.0010)
15%	0.182 (0.0026)	0.091 (0.0027)	0.031 (0.0004)	0.038 (0.0002)	0.065 (0.0022)	0.044 (0.0004)
30%	0.175 (0.0034)	0.083 (0.0015)	0.027 (0.0002)	0.035 (0.0002)	0.051 (0.0020)	0.041 (0.0005)
			10 wake/sleep cycles			
1%	0.197 (0.0013)	0.042 (0.0009)	0.054 (0.0023)	0.034 (0.0008)	0.078 (0.0030)	0.075 (0.0026)
15%	0.183 (0.0019)	0.034 (0.0004)	0.057 (0.0018)	0.030 (0.0001)	0.047 (0.0025)	0.056 (0.0008)
30%	0.176 (0.0025)	0.034 (0.0007)	0.045 (0.0010)	0.030 (0.0002)	0.032 (0.0001)	0.040 (0.0004)
			15 wake/sleep cycles			
1%	0.197 (0.0021)	0.064 (0.0015)	0.066 (0.0023)	0.038 (0.0013)	0.040 (0.0018)	0.060 (0.0019)
15%	0.188 (0.0032)	0.053 (0.0012)	0.049 (0.0011)	0.031 (0.0001)	0.035 (0.0004)	0.046 (0.0002)
30%	0.164 (0.0028)	0.045 (0.0003)	0.052 (0.0016)	0.032 (0.0001)	0.034 (0.0001)	0.051 (0.0004)
			20 wake/sleep cycles			
1%	0.219 (0.0035)	0.051 (0.0013)	0.152 (0.0034)	0.063 (0.0022)	0.073 (0.0055)	0.063 (0.0029)
15%	0.228 (0.0006)	0.043 (0.0002)	0.075 (0.0004)	0.031 (0.0001)	0.029 (0.0001)	0.042 (0.0003)
30%	0.208 (0.0004)	0.034 (0.0000)	0.063 (0.0003)	0.030 (0.0000)	0.029 (0.0001)	0.039 (0.0002)

Table 5. Out of sample MSE: mean values and standard errors. Function 4.

Down	Noise 0.01	Noise 0.1	Noise 1.0	Noise 10.0	Noise 30.0	Noise 50.0
			Standard MLP			
n/a	0.002 (0.0000)	0.002 (0.0001)	0.002 (0.0001)	0.027 (0.0001)	0.066 (0.0003)	0.133 (0.0002)
			5 wake/sleep cycles			
1%	0.005 (0.0001)	0.004 (0.0001)	0.003 (0.0001)	0.037 (0.0011)	0.093 (0.0006)	0.152 (0.0027)
15%	0.013 (0.0001)	0.011 (0.0003)	0.014 (0.0002)	0.044 (0.0009)	0.087 (0.0010)	0.141 (0.0021)
30%	0.028 (0.0005)	0.027 (0.0006)	0.041 (0.0007)	0.065 (0.0018)	0.087 (0.0017)	0.124 (0.0025)
			10 wake/sleep cycles			
1%	0.007 (0.0005)	0.004 (0.0002)	0.010 (0.0011)	0.065 (0.0028)	0.184 (0.0040)	0.156 (0.0038)
15%	0.014 (0.0004)	0.012 (0.0002)	0.010 (0.0003)	0.066 (0.0018)	0.108 (0.0033)	0.093 (0.0016)
30%	0.024 (0.0003)	0.026 (0.0003)	0.032 (0.0007)	0.095 (0.0018)	0.085 (0.0020)	0.083 (0.0028)
			15 wake/sleep cycles			
1%	0.012 (0.0008)	0.006 (0.0008)	0.014 (0.0021)	0.070 (0.0026)	0.147 (0.0041)	0.144 (0.0057)
15%	0.015 (0.0003)	0.013 (0.0009)	0.010 (0.0002)	0.104 (0.0019)	0.073 (0.0026)	0.113 (0.0044)
30%	0.021 (0.0001)	0.024 (0.0002)	0.030 (0.0004)	0.122 (0.0012)	0.062 (0.0008)	0.090 (0.0015)
			20 wake/sleep cycles			
1%	0.008 (0.0003)	0.008 (0.0004)	0.032 (0.0023)	0.061 (0.0027)	0.054 (0.0034)	0.166 (0.0065)
15%	0.015 (0.0003)	0.019 (0.0007)	0.018 (0.0002)	0.082 (0.0013)	0.036 (0.0003)	0.121 (0.0012)
30%	0.022 (0.0001)	0.030 (0.0001)	0.034 (0.0001)	0.098 (0.0008)	0.036 (0.0001)	0.115 (0.0010)

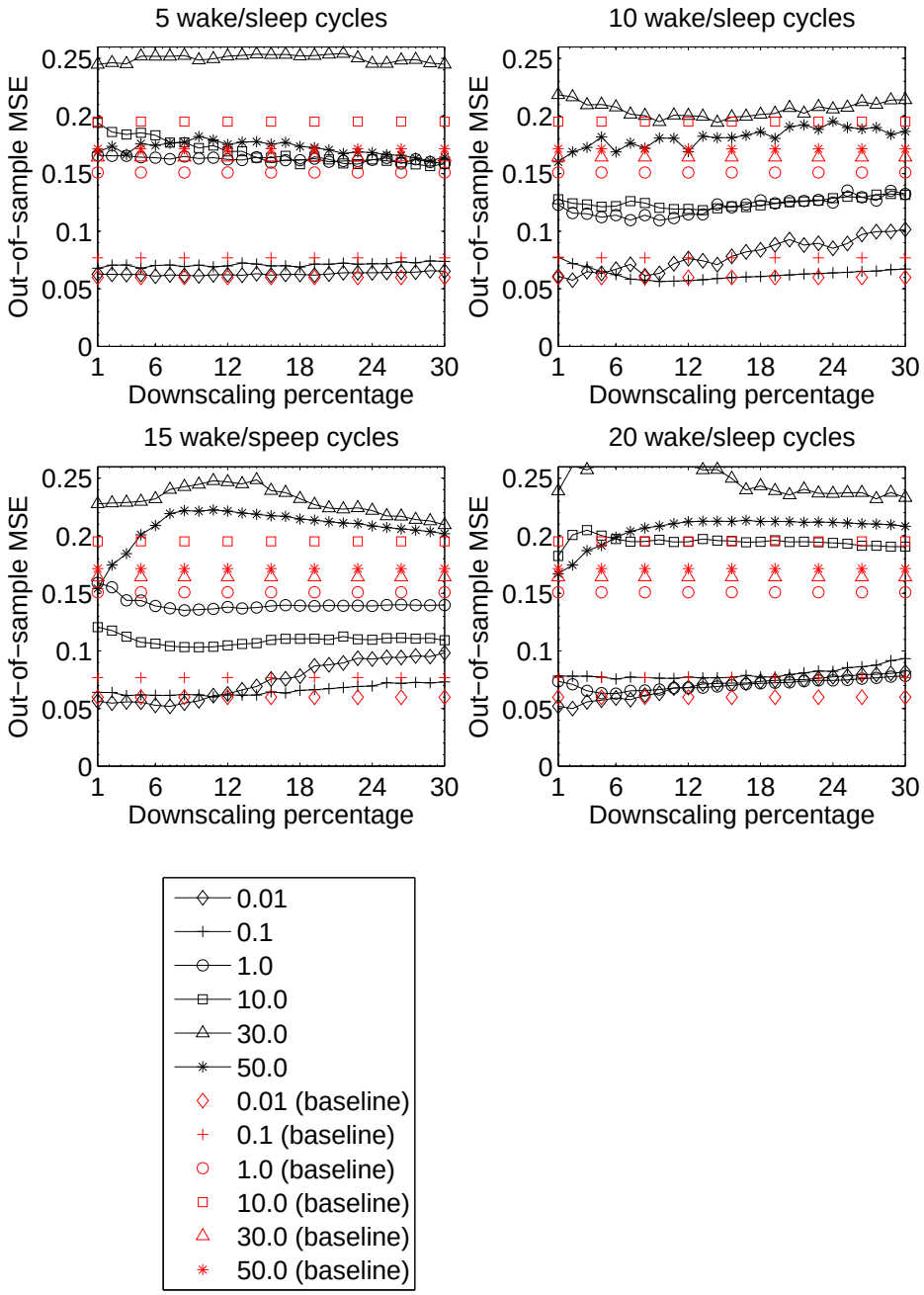


Figure 1. Out-of-sample results for problem F_1 with six different noise levels.

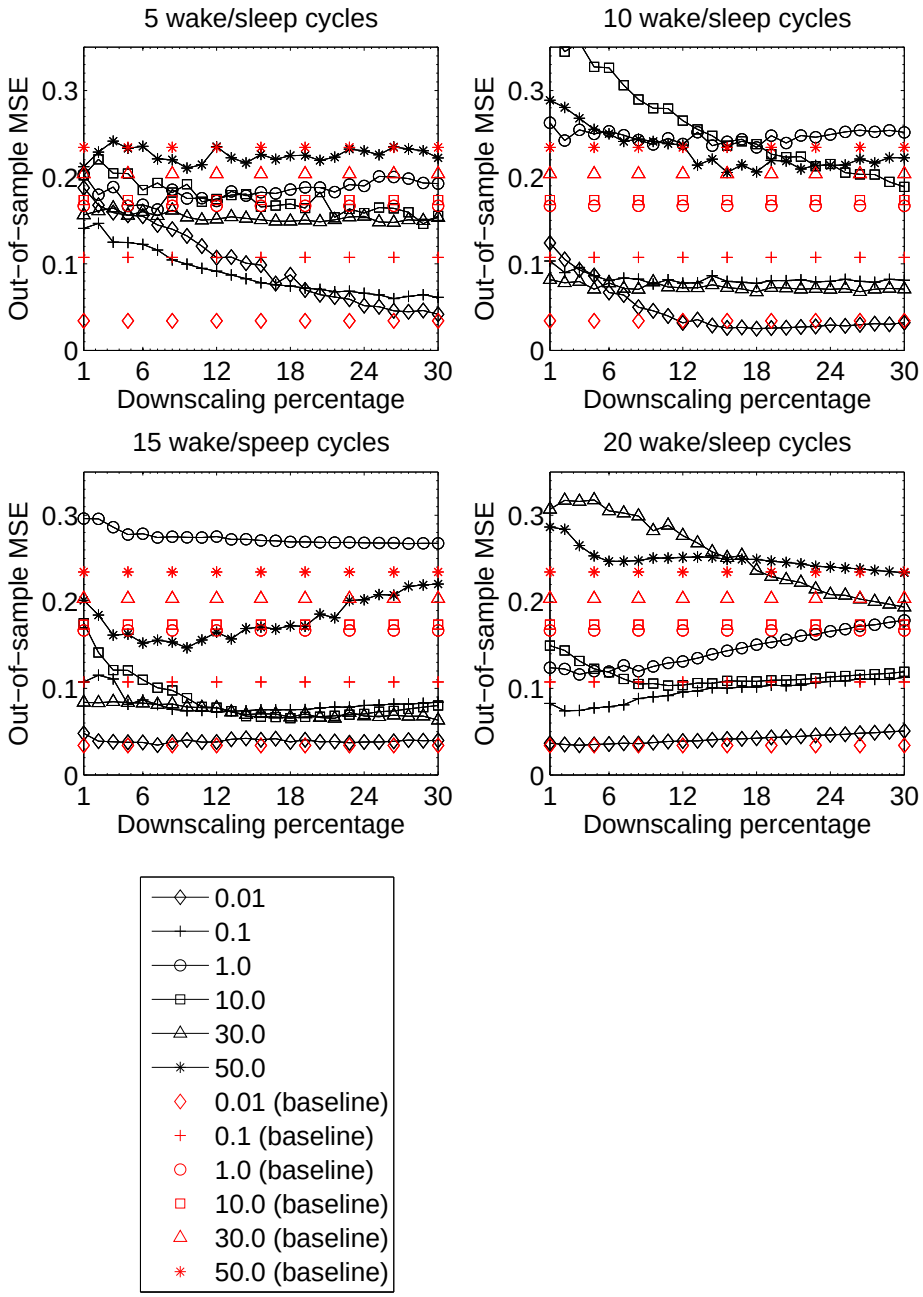


Figure 2. Out-of-sample results for problem F_2 with six different noise levels.

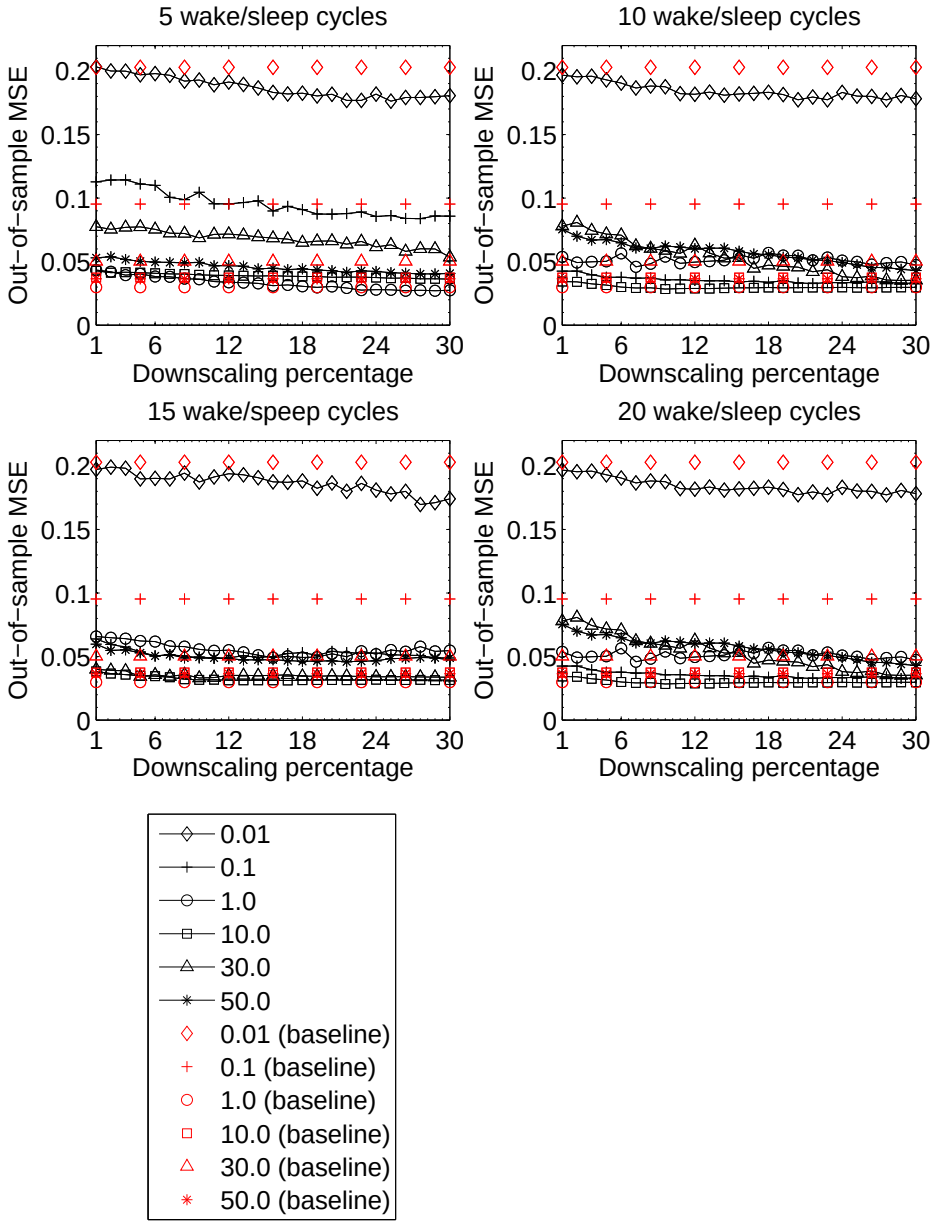


Figure 3. Out-of-sample results for problem F_3 with six different noise levels.

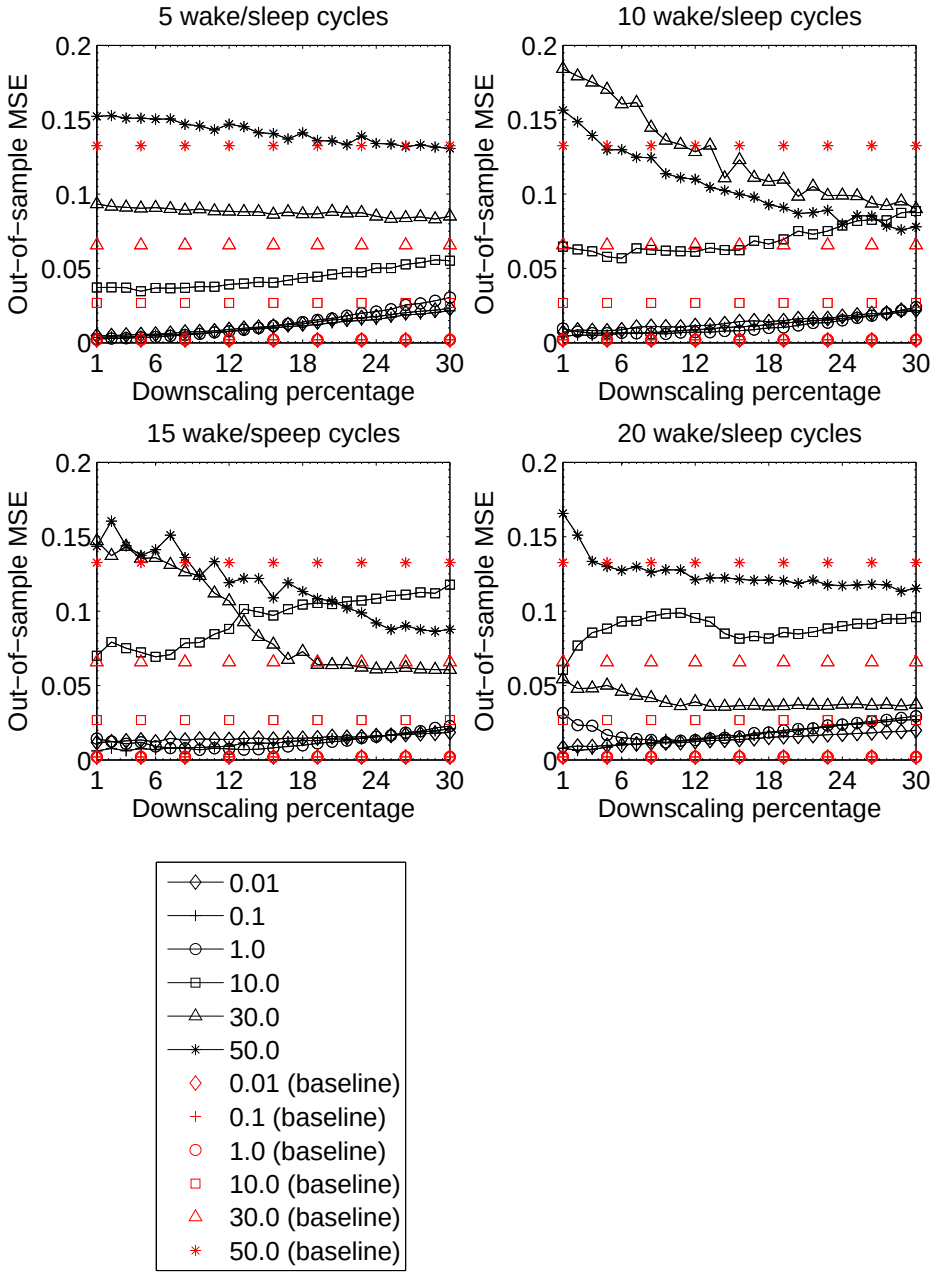


Figure 4. Out-of-sample results for problem F_4 with six different noise levels.

5. Conclusion

In prediction problems, fitting the training data too closely can be counterproductive. Reducing the expected loss on the training data beyond some point causes the population-expected loss to stop decreasing, and often start to increase. Regularisation methods in MLPs, like weight decay, prevent such overfitting by constraining the magnitude of the adaptive weights during the learning phase. In the chapter we showed that simulating a simple weight downscaling mechanism during a sleep phase can similarly to weight decay exert a positive effect on generalisation in the case of noisy datasets.

Controlling the parameters defined as the *downscaling percentage* and the *number of wake/sleep cycles* regulate the degree to which the expected loss on the training data is minimised. Each of the two parameters controls the degree-of-fit and thus affects the best value for the other one. Decreasing the value of downscaling percentage, increases the best value for the wake/sleep cycles. Ideally, one should estimate optimal values for both by minimising a model selection criterion jointly with respect to the values of the two parameters. There are also computational considerations; increasing the value of sleep/wake cycles produces a proportionate increase in the computation. Its value should be made as large as is computationally feasible. The value of downscaling percentage should then be adjusted using cross-validation.

A final observation concerns the nature of the learning process via a number of sleep/wake cycles. Unlike fitting the weights of the network during a number of epochs with a fixed learning rate, the sleep/wake approach instead *learns more slowly*. In general, it has been repeatedly advocated in the statistical machine learning literature that learning methods that learn slowly tend to generalise well.

Acknowledgement

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant Number 08/SRC/FM1389.

References

[1] Axmacher, N., Draguhn, A., Elger, C., Fell, J.: Memory processes during sleep: beyond the standard consolidation theory. Cellular and Molecular Life Sciences, **66**:2285–2297 (2009)

[2] Bpurce, J. and Harris, K. (2007). Do thin spines learn to be mushroom spines that remember?, Current Opinion Neurobiology, 17:381–386.

[3] Cirelli, C., Tononi, G.: Is Sleep Essential. PLoS Biology, **6(8)**:e216 (2008)

[4] Cirelli, C.: The genetic and molecular regulation of sleep: from fruit flies to humans,. Nature Reviews Neuroscience, **10**:549–560 (2009)

[5] Diekelmann, S., Born, J.: The memory function of sleep. Nature Reviews Neuroscience, **11**:114–126 (2010)

[6] Euston, D., Steenland, H.: Memories-getting wired during sleep. Science, **344(6188)**:1087–1088 (2014)

[7] Greenspan, R., Tononi, G., Cirelli, C., Shaw, P.: Sleep and the fruit fly. Trends in Neurosciences, **24(3)**:142–145 (2001)

[8] Hardt, O., Nader, K., Nadel, L.: Decay happens: the role of active forgetting in memory. Trends in Cognitive Sciences, **17(3)**:111–120 (2013)

[9] Hinton, G., Dayan, P., Frey, B., Neal, R.: The wake-sleep algorithm for unsupervised neural networks. Science, **268(5214)**:1158–1161 (1995)

[10] james, W. (1890). Principles of Psychology, Holt, New York.

[11] Karnin, E.: A Simple Procedure for Pruning Back-Propagation Trained Neural Networks. IEEE Transactions on Neural Networks, **1(2)**:239–242 (1990)

[12] Krogh, A., Hertz, J.: A Simple Weight Decay Can Improve Generalization. In: Advances in Neural Information Processing Systems 4, pp. 950–957, Morgan Kaufmann, San Mateo (1995)

[13] LeCun, Y., Denker, J., Solla, S., Howard, R., Jackel, L.: Optimal brain damage. In: Advances in Neural Information Processing Systems (NIPS), vol. 2, pp. 598–605, Morgan Kaufmann, San Mateo (1989)

[14] Lima, S., Rattenborg, N., Lesku, J., Amlaner, C.: Sleeping under the risk of predation. Animal Behaviour, **70**:723–736 (2005)

[15] Majumdar A, Cesario WC, White-Grindley E, Jiang H, Ren F, Khan M, Li, L, Choi E, Kannan K, Guo F, Unruh J, Slaughter B, Si K. (2012) Critical role of amyloid-like oligomers of Drosophila Orb2 in the persistence of memory. Cell 148: 515529. doi:10.1016/j.cell.2012.01.004

[16] Marr, D. (1971). Simple memory: a theory for archicortex, Philosophical Transactions of the Royal Society of London, Series B, 262:23–81.

[17] McGaugh, J.: Memory–a Century of Consolidation. Science, **287(5451)**: 248–251 (2000)

[18] Müller, G., Pilzecker, A.: Experimentelle Beiträge zur Lehre vom Gedächtniss. Zeitschrift für Psychologie. Ergänzungsband (1900)

[19] Nadel, L. and Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex, Current Opinion in Neurobiology, 7:217–227.

[20] Plaut, D., Nowlan, S., Hinton, G.: Experiments on Learning by Back Propagation. Technical Report Computer Science Department (CMU-CS-86-126), Carnegie-Mellon University, Pittsburgh (1986)

[21] Rattenborg, N., Amlaner, C., Lima, S.: Behavioral, neurophysiological and evolutionary perspectives on unihemispheric sleep. Neuroscience and Biobehavioral Reviews, **24**:817–842 (2000)

[22] Rattenborg, N., Martinez-Gonzalez, D., Lesku, J.: Avian sleep homeostasis: Convergent evolution of complex brains, cognition and sleep functions in mammals and birds. Neuroscience and Biobehavioral Reviews, **33**:253–270 (2009)

[23] Rudy, J. (2014). The Neurobiology of learning and memory (2nd ed), Sinauer Associates Inc., Sunderland Massachusetts

[24] Schulz, H.: Rethinking sleep analysis. Comment on the AASM (American Academy of Sleep Medicine) Manual for the Scoring of Sleep and Associated Events. Journal of Clinical Sleep Medicine, **4(2)**: 99–103 (2008)

[25] Sejnowski, T., Destexhe, A.: Why do we sleep? Brain Research, **886**:208-223 (2000)

[26] Siegel, J.: Do all animals sleep? Trends in Neuroscience, **31(4)**:208–213 (2008)

[27] Siegel, J.: Sleep viewed as a state of adaptive inactivity. Nature Reviews Neuroscience, **10**:747–753 (2009)

[28] Sietsma, J., Dow, R.: Neural net pruning-why and how. In: Proceedings of 1988 IEEE International Conference on Neural Networks, **1**:325–333, IEEE Press (1989)

[29] Sietsma, J., Dow, R.: Creating artificial neural networks that generalise. Neural Networks, **4(1)**:67–79 (1991)

[30] Squire, L. (1987). Memory and the Brain, Oxford University Press.

[31] Stickgold, R.: Sleep: off-line memory reprocessing. Trends in Cognitive Science, **2(12)**:484–492 (1998)

[32] Tononi, G., Cirelli, C.: Sleep and synaptic homeostasis: a hypothesis. Brain Res Bull. **62(2)**:143–150 (2003)

[33] Tononi, G., Cirelli, C.: Sleep function and synaptic homeostasis. Sleep Med Rev. **10(1)**:49–62 (2006)

[34] Tononi, G., Cirelli, C.: Perchance to Prune. Scientific American, **309**:34–39 (2013)

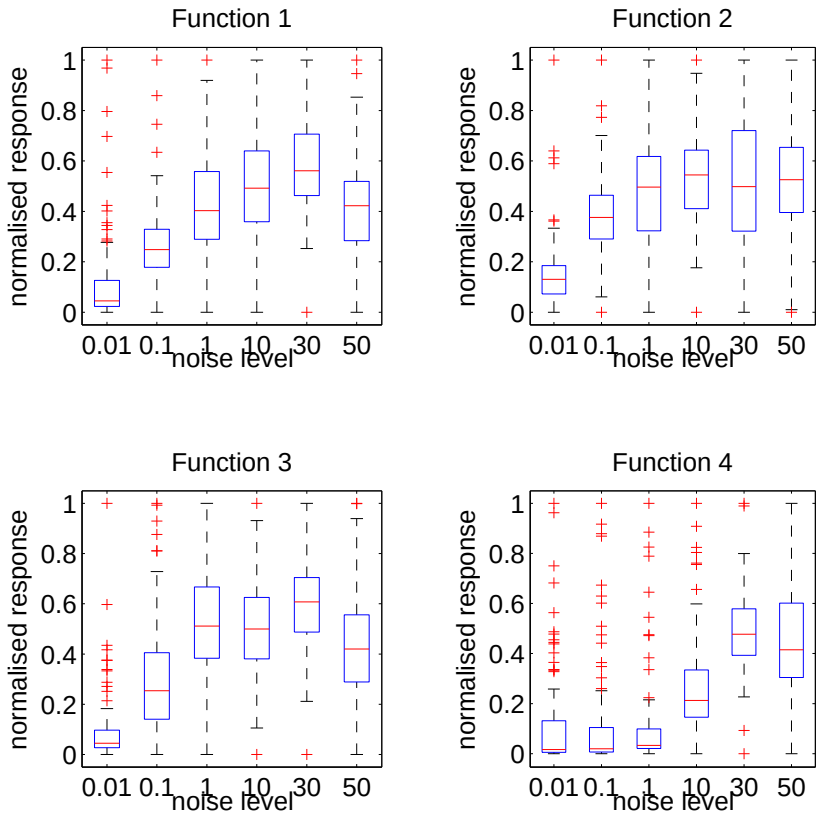
[35] Teyler, T. and DiScenna, P. (1986). The hippocampal memory indexing theory, Behavioral Neuroscience, 100:147–152.

[36] Walker, M., Stickgold, R., Alsop, D., Gaab, N., Schlaug, G.: Sleep-dependent motor memory plasticity in the human brain. Neuroscience, **133(4)**:911-917 (2005)

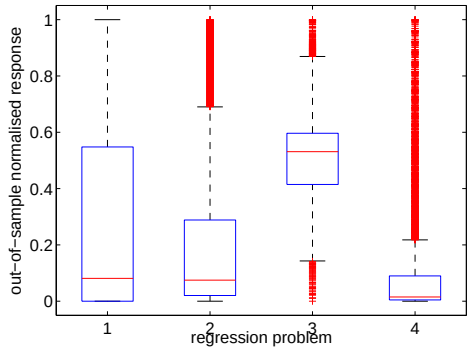
[37] Weigend, A., Rumelhart, D., Huberman, B.: Generalization by weight-elimination with application to forecasting. In: R. P. Lippmann, J. Moody, & D. S. Touretzky (eds.), Advances in Neural Information Processing Systems 3, pp. 875–882, San Mateo, CA, Morgan Kaufmann (1991)

[38] White-Grindley E, Li L, Mohammad Khan R, Ren F, Saraf A, Florens L, Si K (2014) Contribution of Orb2A Stability in Regulated Amyloid-Like Oligomerization of Drosophila Orb2. PLoS Biol 12(2): e1001786. doi:10.1371/journal.pbio.1001786

[39] Yang, G., Lai, C. S., W., Cichon, J., Ma, L., Li, W., Gan, W-B.: Sleep promotes branch-specific formation of dendritic spines after learning. Science, **344(6188)**:1173–1078 (2014)



(a)



(b)

Figure 5. (a) Histogram of in-sample normalised values of the response variable for different regression problems of Table 1. (b) Histogram of out-of-sample normalised values of the response variable for different regression problems of Table 1.

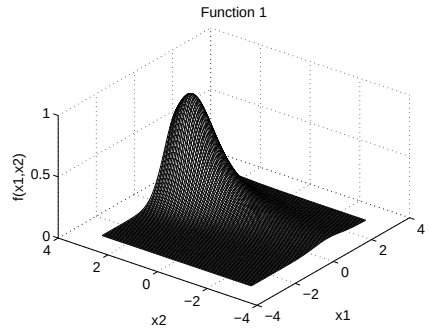


Figure 6. Plot of regression problem 1 of Table 1.

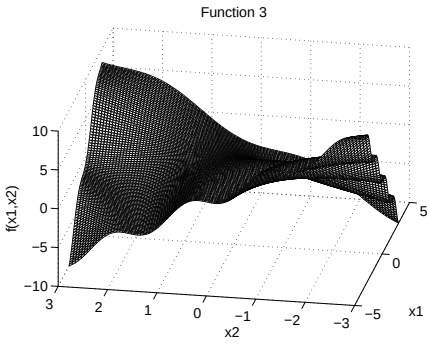


Figure 7. Plot of regression problem 3 of Table 1.

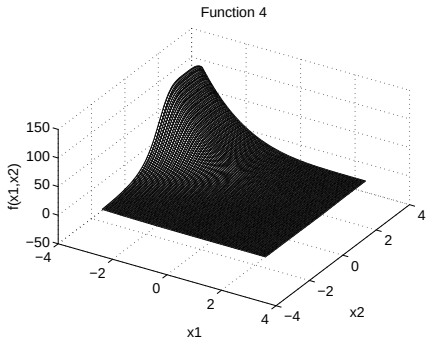


Figure 8. Plot of regression problem 4 of Table 1.