SHORT-TERM SYNAPTIC PLASTICITY UNDERLIES ACTIVITY-SILENT WORKING MEMORY: A FUNCTIONAL SPIKING NEURON MODEL

Bachelor’s Project Thesis
Matthijs Pals, s2749440, m.pals.3@student.rug.nl
Supervisor: Dr J.P. Borst

Abstract: We developed an end-to-end model of working memory (WM) based on the latest theories regarding the maintenance of information in the human brain. While it used to be widely assumed that storage in WM is maintained via persistent recurrent activity, recent studies showed that information can be maintained without persistent firing; there is storage in so called activity-silent states. A candidate mechanism underlying this type of storage is short-term synaptic plasticity (STSP), where the strength of connections between neurons rapidly changes to reflect new information being loaded in WM. In order to demonstrate that STSP by means of calcium-mediated synaptic facilitation can lead to behaviour similar to humans, STSP was integrated in a large-scale functioning spiking neuron model. This model was able to execute a delayed-response task in which a randomly oriented grating had to be maintained in WM. An earlier study measured the neural activity of human participants during this task. It was demonstrated that displaying a task-irrelevant stimulus during the maintenance period can reveal what is stored in activity-silent states. In support of our model’s plausibility, we showed that both its performance and neural activity during the task correspond to the human data of that study. We conclude that information in WM can be effectively maintained in activity-silent states by means of calcium-mediated STSP.

1. Introduction
The ability to temporarily hold visual information in Working Memory (WM) is a crucial part of day-to-day life; it is what allows us to locate a book in a stocked shelf, a familiar face in a bustling city square and our keys on a cluttered desk (e.g., Baddeley & Hitch, 1974; Barak & Tsodyks, 2014). The maintenance of visual information in WM is often studied by means of a delayed-response task (Fuster, 2015). During this task a memory item is briefly presented, followed by a delay period. The delay period ends with the presentation of a probe which the participants compare need to with the memory item held in WM in order to give the correct response.

The maintenance of information during the delay period of such tasks was thought to be mediated by ensembles of continuously spiking neurons, starting with the influential work of Fuster and Alexander (1971) and later followed by many others (e.g. Goldman-Rakic, 1995). While there is no denying that spiking activity indeed plays a crucial role in maintaining information in WM, spiking activity during delay periods appears to be not always as persistent as was often believed (Lundqvist, Herman, & Miller, 2018). Earlier studies might have relied too much on averaging spiking across trials, creating the impression that spiking is persistent even though it might consist of sparse bursts (Lundqvist et al., 2016).

Recently it was shown that neural activity related to maintenance of a stimulus quickly disappears after stimulus presentation (Myers et al., 2015; Wolff, Jochim, Akyürek, & Stokes, 2017). However despite the lack of persistent activity, information can still be maintained. It appears that neural activity is reserved only for the most relevant time points, for instance around the presentation of a probe or memory item in a delayed-response task (Barak, Tsodyks, & Romo, 2010; Stokes, 2015). In between those time points information seems to be stored in so called activity-silent states.

One of the candidate mechanisms that mediates storage in activity-silent states is short-term synaptic plasticity (STSP), where the strength of connections between neurons rapidly
changes to reflect new information being loaded in the network (Mongillo, Barak, & Tsodyks, 2008). Indeed, it was previously shown that synapses in areas implicated in WM can be facilitating (Tsodyks, Pawelzik, & Markram, 1998; Wang et al., 2006), possibly as a consequence of residual calcium building up in the presynaptic terminals (Jackman & Regehr, 2017; Zucker & Regehr, 2002). In this way residual calcium effectively leaves a ‘synaptic trace’ of what is currently stored in WM.

An elegant implementation of activity-silent storage by means of STSP through calcium kinetics was proposed by Mongillo and colleagues (2008). They developed a model that showed that information can effectively be maintained by calcium-mediated synaptic facilitation in recurrent networks of simulated spiking neurons. After providing an input to the network, the connections from neurons that are firing will be facilitated. Subsequently, stored information can be read out by applying a network-wide non-specific input that will be mostly subthreshold for non-facilitated neurons but leads to firing of facilitated neurons.

In this study we integrated calcium-mediated STSP as developed by Mongillo and colleagues (2008), in a functioning spiking neuron model of WM that can execute a delayed-response task. Decision-making will be guided by a match-filter process (Myers et al., 2015; Stokes, 2015), transforming the input to task-appropriate output. The model will be implemented using Nengo, a Python library for simulating large-scale neural models with a clear link between spiking activity and representation (Bekolay et al., 2014; Eliasmith, 2013; Eliasmith et al., 2012).

Thus we aim to show that a model implementing the mechanism proposed by Mongillo and colleagues (2008) in biologically plausible neurons can lead not just to efficient and robust storage, but also to function. To evaluate the model, we will compare its accuracy on a delayed-response task to the accuracy from human participants on this task by Wolff and colleagues (2017). Furthermore, to evaluate the plausibility of our model, we will also compare its spiking activity during the task to the electroencephalography (EEG) data from Wolff and colleagues (2017), who recently developed an innovative method, - a ‘neural sonar’ – to probe activity-silent brain states: when the WM network is ‘pinged’ by a task-irrelevant stimulus, ensuing neural activity reveals what is currently maintained in activity-silent states.

In the remainder of the introduction, we will explain the ‘neural sonar’ method to probe activity-silent states. We will also briefly introduce Nengo, the used neural architecture in more detail. This section will be followed by a section containing a detailed description of the model and methods and a section that includes a description of our model’s behaviour. Finally we will conclude with a discussion of our results and compare our model to previously proposed models.

1.1. Probing hidden working memory states

Neuroimaging data can provide us with a powerful tool to constrain computational cognitive models (Borst, Nijboer, Taatgen, Van Rijn, & Anderson, 2015). Given that we traditionally rely on non-invasive techniques such as electroencephalography (EEG) and functional magnetic resonance imaging (fMRI), it might at first appear problematic to evaluate a model of activity-silent storage. There is little measurable activity to constrain our model with when information is maintained in an activity-silent state. However, recently a clever method to probe these ‘hidden’ states by means of EEG was developed (Wolff, Ding, Myers, & Stokes, 2015; Wolff et al., 2017).

Wolff and colleagues (2017) performed a retro-cue delayed-response task. See Figure 1.1 for an overview of a trial. Each trial started with the display of two memory items (randomly oriented grating stimuli, Figure 2.3) followed by a cue indicating which of the two presented stimuli was to be maintained in memory. After a delay-period a new grating stimuli was presented and participants were to indicate

![Figure 1.1: The retro-cue delayed-response task performed by the participants of Wolff and colleagues (2017).](image-url)
whether this was rotated clockwise or anticlockwise with respect to the cued memory item.

In order to probe what is maintained in activity-silent states, Wolff and colleagues (2017) inserted a constant high-contrast task-irrelevant stimulus, coined an ‘impulse’, during the delay part of the trial. They showed that it is possible to decode the orientation of the cued memory item from the EEG activity in reaction to the presentation of the ‘impulse’. In other words, when the WM network is ‘pinged’ by a task-irrelevant stimulus, the ensuing signal allows for the decoding of what is currently maintained in activity-silent states. To clarify, an analogy with a sonar is made; where a sonar probes hidden contours of unseen structures on the seabed the ‘impulse’ probes the hidden structure of the WM network.

1.2. Neural network architecture

In order to implement a functional spiking-neuron model of WM we used Nengo. Nengo is a python library that allows one to build large-scale brain models that link single cell activity to demonstrative cognitive abilities (Bekolay et al., 2014). It has been used to create detailed cognitive models (eg., Choo & Eliasmith, 2010; Singh & Eliasmith, 2006) and has been successfully used to create Spaun, a functioning brain model consisting of 2.5 million neurons (Eliasmith et al., 2012). Nengo makes use of a theoretical framework called the Neural Engineering Framework (NEF; Eliasmith & Anderson, 2003). Two relevant principles Nengo models and the NEF are based are representation and transformation.

**Representation:** Information is represented as a vector of real numbers that can be encoded and decoded from the collective spiking activity of a population of neurons, called an ensemble. Encoding is mediated by giving each neuron a non-linear tuning curve that characterizes their general response to the incoming signal. Decoding is a linear process: the activity of each neuron in an ensemble is weighted by a constant (the decoder) and summed up over time in order to decode the represented vector.

**Transformation:** Connections between ensembles of neurons allow for the transformation of the vectors represented by these ensembles by specifying a function. Here the NEF acts as a ‘neural compiler’: given a desired function, the connection weight matrix between ensembles is calculated so that this function is approximated. Besides pre-calculating a weight matrix, plasticity can be introduced by making use of biologically plausible learning rules, both supervised and unsupervised (Bekolay, Kolbeck, & Eliasmith, 2013).

2. Methods

2.1. Implementation of short-term synaptic plasticity

To account for short-term synaptic plasticity, we integrated the calcium kinetics mechanism of Mongillo et and colleagues. (2008) in Nengo (Figure 2.1). Both the Nengo implementation of STSP and the model are available (https://github.com/Matthijspals/STSP). Synaptic efficiency was based on two parameters: the amount of available resources to the presynaptic neuron ($x$, normalised to be between 0 and 1) and the fraction of resources used each time a neuron fires ($u$), reflecting the residual presynaptic calcium level.

Spiking leaky integrate-and-fire (LIF) neurons are computational efficient while retaining a degree of biological plausibility (eg., Burkitt, 2006).

![Figure 2.1: Nengo reproduction of Mongillo and colleagues (2008). Top: spikes in presynaptic neuron. Middle: calcium ($u$) and resources ($x$) of presynaptic neuron, $u$ increases and $x$ decreases when the presynaptic neuron spikes. Bottom: resulting postsynaptic voltage, note the synaptic depression after the first spike train and the synaptic facilitation during the later spike.](https://github.com/Matthijspals/STSP)
We therefore extended the current Nengo implementation of LIF neurons: for all LIF neurons to which we want to apply STSP, every simulation time step $u$ and $x$ are calculated according to equation 2.1 and 2.2, respectively. When a neuron fires, its resources $x$ are decreased by $ux$, mimicking neurotransmitter depletion. At the same time, its calcium level $u$ is increased, mimicking calcium influx into the presynaptic terminal. Both $u$ and $x$ relax back to baseline with time constants $\tau_D$ (0.2s) and $\tau_F$ (1.5s), respectively. The mechanisms are described by:

\[
\begin{align*}
\frac{dx}{dt} &= \frac{1-x}{\tau_D} - u x \delta(t - t_{sp}) \\
\frac{du}{dt} &= \frac{u - u(t - t_{sp})}{\tau_F} - U (1 - u) \delta(t - t_{sp})
\end{align*}
\]  

Where $x$ represents the available resources, $u$ represents the residual calcium level and $U$ its baseline level, $\tau_F$ is the facilitating time constant and $\tau_D$ the depressing time constant, $\delta$ represents the Dirac delta function, $t$ the simulation time and $t_{sp}$ the time of a presynaptic spike.

Outgoing connection weights of neurons implementing STSP are determined by both their initial connection weight and their current synaptic efficiency. Initial connections weights are calculated by the NEF, while synaptic efficiency is set to the product of the current value of $u$ and $x$ of the presynaptic neuron, normalised by their baseline value (equation 2.3). This results in a system where after a neuron fires its outgoing connections will be depressed on the time scale of $\tau_D$ and facilitated on the timescale of $\tau_F$ as illustrated in Figure 2.1.

\[
\frac{dw_{ij}}{dt} = \frac{e u}{c} w_{ij}^{0}
\]

Where $w_{ij}$ represents the connection weight between neuron $i$ and $j$ and $w_{ij}^{0}$ the initial connection weight between neuron $i$ and $j$.

2.2. The Model

2.2.1 Model architecture

The short-term synaptic plasticity mechanism as described above was integrated in a functioning spiking neuron model. The overall architecture of the model can be seen in Figure 2.2. As described above, Wolff and colleagues (2017) performed a delayed-response task. They showed significant lateralization of EEG activity measured with posterior electrodes after presentation of the cue. Based on this observation we hypothesize that distinct populations of neurons are responsible for maintaining visual stimuli presented in the left and right visual field. Correspondingly, we divided our model into two relatively independent modules, each responsible for one of two incoming stimuli.

Both modules consist of four ensembles of 2000 LIF neurons. In short, visual input enters via a sensory ensemble that transforms it into a 24 dimensional vector. This vector is then sent to a recurrently connected memory ensemble exhibiting STSP. Decision is guided by an ensemble that receives input from both layers.

2.2.2 Sensory ensemble

In order to demonstrate that our model is able to deal with real-world input, it will be shown stimuli similar to those shown to the participants in the study by Wolff and colleagues (2017). To describe the relationship between representation and real-world stimuli it can be assumed that the brain makes use of a statistical model, not unlikely a parametrized model, where a small number of parameters capture the overall shape of the data (Eliasmith, 2013). To find such a model we need a set of basis functions that will be good at both describing incoming images and the encoders of the neurons receiving these
images. These basis functions can be found by applying singular value decomposition (SVD) to a matrix containing both the images and the encoders. The images consisted of the stimuli in the experiment, while the encoders were two-dimensional Gabor filters, defined by a sinusoidal plane wave multiplied by a Gaussian function. Gabor filters have previously been shown to accurately describe the response profile of simple cells in the cat (Jones & Palmer, 1987) and macaque (Ringach, 2002) striate cortex and seem to underlie early stages of visual processing. Thus the SVD mediates a biologically plausible method that results in stimuli being represented by 24-dimensional vectors which will be conveyed to the memory ensemble.

2.2.2 Memory ensemble

Previous WM models assumed WM networks to be highly recurrent (e.g., Camperi & Wang, 1998; Miller, Brody, Romo, & Wang, 2003). Furthermore, it was previously shown that areas implicated in WM contain recurrent connections that are highly facilitated on a short time-scale (Wang et al., 2006). Thus, we assumed the memory ensemble to have recurrent connections exhibiting STSP as described above. The first stimulus during a trial will drive facilitation of the neurons representing this stimulus. Subsequently, the neural activity resulting from a second stimulus will be guided by this change in connectivity. A non-specific ensemble-wide input should be able to re-activate a stimulus maintained in an activity-silent state, if this input is mostly above threshold for firing for facilitated neurons and subthreshold for non-facilitated neurons (Mongillo et al., 2008). This non-specific input would then also help maintain the stimulus for a longer time period, as reactivation of facilitated synapses will lead to re-facilitation of those connections.

2.2.3 Decision ensemble

In order to give our mode the ability to produce task-relevant output, an ensemble was needed that could decode the orientation of the stimuli maintained in the WM network. The sensory ensemble represents each stimulus as a distinct vector. The vectors by which each stimuli will be represented were procured by making use of the basis vectors derived by SVD. Next we specified for each possible represented stimuli what the corresponding decoded orientation should be. This information can then be used to define a function that takes the 24 dimensional vectors represented by the memory and sensory ensembles as input and gives the corresponding orientations as output. The NEF yields the connection matrix at the neural level that approximates this function.

The stimulus orientation was not directly decoded as the angle theta, but rather by the sine and cosine of theta. Thus this ensemble represents four dimensions, sine and cosine of both the sensory and the memory ensemble. Decoding sine and cosine of theta is robust, as the ratio between the two determines the stimulus orientation, while that is not the case when decoding theta directly. Furthermore, the symmetry of the sine functions provides a natural solution for the symmetric nature of the stimuli, as during this experiment a stimulus with an orientation of -90° contains exactly the same pixels as a stimulus with an orientation of 90° degrees and was therefore assumed to result in the same neural activity.

2.2.4 Decision processing and output ensemble

This resulted in a system where the decision ensemble received the sine and cosine of the orientation of the stimuli currently represented by the sensory and memory ensembles. When a probe stimulus is presented, the orientation received from the sensory ensemble was driven completely by the incoming stimulus. The stimulus orientation of the memory ensemble however, was driven by a dynamic combination of activity resulting from the incoming stimulus and activity resulting from facilitated connections. In other words, the orientation represented by the memory population reflected the orientation of the probe stimulus ‘tuned’ by the orientation of the memory item stored in facilitated synapses, over time reverting to the new probe stimulus. To estimate the orientation difference between the memory item and the probe, the outgoing connections from the decision layer subtract the two represented orientations. Thus the resulting, one dimensional value, represented by the output ensemble, reflects the signed difference between the
orientation of the memory item and orientation of the probe stimulus.

2.3. The Experiment

2.3.1 Stimuli

Input to the model consisted of two images of 128 by 128 pixels. Stimuli were generated using Psychopy, an open-source Python application (Peirce et al., 2019). Stimuli consisted of a circle on a grey background (RGB = 128, 128, 128). Memory items and probe stimuli were sine-wave gratings with a diameter of 128 pixels and spatial frequency of 0.034 cycles per pixel. The phase was randomized within and across trials. Examples can be seen in Figure 2.3.

For each trial, the orientation of the memory items was randomly selected from a uniform distribution of orientations. The impulse stimulus consisted of a ‘bull’s-eye’ stimulus of the same size and spatial frequency as the memory items. It was presented at twice the contrast compared to the grating stimuli. The angular differences between the memory item and the corresponding probe stimulus were uniformly distributed across seven angle differences (3°, 7°, 12°, 18°, 25°, 33°, 42°).

2.3.2 Procedure

An overview of the experimental procedure can be seen in Figure 2.4. The model completed the retro-cue delayed-response task from Wolff and colleagues (2017). Each trial started with two memory items, consisting of two randomly oriented low-contrast gratings, that are each presented to the sensory ensemble of one of the two modules for 250 ms. After 800 ms, Wolff and colleagues (2017) presented a retro-que that signalled which of the two previously presented items would be tested. The offset of the retro-que was followed by significant lateralisation in posterior electrodes. To mimic this the memory ensemble of the cued module is shortly reactivated by means of a non-specific ensemble wide input (cf. Mongillo et al., 2008) 1150 ms after offset of the memory item. The impulse stimulus was presented to the sensory ensembles of both sub-models for 100 ms, 1700 ms after the offset of memory items. After another delay of 400 ms, the memory probe was presented to the sensory ensembles for 250 ms. To mimic different participants in the experiment, every 1344 trials the model seed was reset and new random Gabor filters were generated to use as encoders for the sensory ensembles. In total the model performed 30 sets of 1344 trials.

To decode a response from the model, and determine the difference between the memory item and probe stimulus, the activation of the output ensemble was integrated for 50 ms, 30 ms after presentation of the probe. A suitable period of integration was determined by means of visual inspection of the neural activation of the output ensemble and kept constant among all trials. A positive result after integration was assumed to correlate to a response stating that the probe stimulus was rotated clockwise relative to the memory item, while a negative result was related to a counter-clockwise difference in orientation. Integrating the evidence corresponding to two distinct decisions has been widely used before in accumulator models of perceptual decision making (Palmeri, Schall, & Logan, 2015).

2.4. Analysis

Wolff and colleagues (2017) decoded the orientation of the maintained stimuli in the WM network from the activity of EEG electrodes. They showed that decoding accuracy quickly dropped after presentation of the memory item, but returned again during presentation of subsequent stimuli, indicating that in-between stimuli information is maintained in activity-
silent states. Correspondingly we examined the spiking activity and representational quality of our model during the task, in order to evaluate if it indeed successfully maintains information without persistent firing.

Furthermore Wolff and colleagues (2017) showed that neural activity as a result from the impulse stimulus reveals the orientation only of the queued memory item. This indicates that a stimulus is only maintained in activity-silent states if it is still needed for the task, otherwise it is quickly cleared from the network. To see if our model has both the same storage and forgetting capabilities we examined the vectors represented by the memory ensembles of both the cued and uncued module during presentation of the impulse.

Finally, in order to see if our model not only corresponds to neural activity, but also gives rise to function similar to human participants, we compared the accuracy of the model to accuracy of the human participants on the task for various orientation differences between probe and memory item.

3. Results

In order to validate that any maintenance of information in our model is realised in activity-silent states and not by persistent activity, we inspected the spiking activity of the WM network. Figure 3.1 (top) shows the spiking activity of the neurons in the memory ensembles of both modules during one trial, together with the mean amount of resources ($x$) and calcium ($u$) of the cells in these ensembles. It can be observed that there is activity during and shortly after presentation of stimuli. Furthermore, in the cued module, there is activity after the onset of non-specific ‘top-down’ reactivation mimicking the retro-cue of the original experiment.

We can see that when there is spiking activity the mean value of $x$ and $u$ decrease and increase, respectively. The resulting short-term depression on the time scale of $\tau_D$ can also be observed as the amount of spiking is highest during the onset of a stimulus and then quickly declines.

Next we analysed the vectors represented by the memory ensembles of both the cued and uncued module. As noted in the methods section each stimuli will be represented by the network as a 24-dimensional vector. We were able to derive this vector for each stimuli by using the basis vectors of the SVD used to generate encoders. This allowed us to compare the ‘ideal’ vector of a stimuli, with the vector represented by an ensemble at any time point. A common

![Figure 3.1](image-url)  
**Figure 3.1**: Top: spiking activity of the memory ensembles of both modules together with the mean amount of resources ($x$) and calcium ($u$) in these ensembles during a trial. Bottom: mean cosine similarity between the vector represented by the memory ensembles and the ‘ideal’ vector for various stimuli over 100 trials. These trials contained memory items with a rotation of $0^\circ$ and probes with a rotation of $42^\circ$. 
method to compare the similarity between two vectors is by calculating their cosine similarity, defined by their dot divided by the product of their magnitude.

Figure 3.1 (bottom) shows the absolute cosine similarity between the vector represented by the memory ensemble and the ‘ideal’ vector of the impulse and other stimuli with various orientations. To clearly show the difference between the two modules, the mean cosine similarity was calculated over 100 trials in which both modules were presented with the same memory item and probe, with a rotation of 0° and 42°, respectively. Here we were solely interested in the effect of orientation on representation within a trial, thus these trials were run with a constant within-trial phase (but a varying between-trial phase).

It can be seen in Figure 3.1 that during presentation of the initial stimulus the vector represented by both modules is similar to the ‘ideal’ vector of this stimulus. Moreover, we can see that an increasing orientation difference between the represented vector and an ‘ideal’ vector entails a decreasing cosine similarity between the two. This indicates that similar incoming stimuli were represented by similar vectors and thus by similar firing patterns.

It can also be noted that during the delay-periods there is no firing and thus we could not simply decode what is being represented by the ensembles. We were however able to decode a represented vector briefly after onset of the non-specific reactivation in the cued module; then there was a brief burst of action potentials. It appears that indeed mostly neurons with facilitated connections fire, as the represented vector is similar to the ‘ideal’ vector of the memory item shown at the start of the trial.

In Figure 3.2 we take a closer look at the cosine similarity after onset of the impulse stimulus. The vector represented by the memory ensembles after presentation of the impulse becomes quite similar to the ‘ideal’ vector of the memory item shown at the start of the trial for both the cued and uncued module; it can be seen that when probed with the impulse stimulus, the memory ensemble ‘rings’ back with the vector representing the memory item maintained in facilitated connections. However, only for the cued module does the represented vector become more similar to the ‘ideal’ vector of the memory item than to ‘ideal’ vector of the impulse, offering a possible explanation of why Wolff and colleagues (2017) were able to decode the cued stimulus (but not the uncued stimulus) from the activity after onset of the impulse.

To see if what is maintained in facilitated synapses can be used to produce a relevant response we analysed the value represented by the output ensemble. The output ensembles receives a signal from the decision ensembles and represents a measure of difference between the orientation decoded from the sensory and memory ensembles. Figure 3.3 shows the represented value for the possible orientation differences between (varying) memory items and probes, averaged over all trials. As can be seen in Figure 3.3, both the sign and magnitude of the orientation was represented in the signal.

Figure 3.2: Cosine similarity between the vector in the memory ensembles and the ‘ideal’ vector of various stimuli, after onset of the impulse. The memory item presented before the impulse had a rotation of 0°.

Figure 3.3: Value represented by the output ensemble after onset of the probe stimulus, averaged separately for each orientation difference over all trials.
Finally we show that activity-silent storage can lead to behaviour similar to what is observed in humans. A natural way to decode a response from the model followed from figure 3.3: we integrated the activation of the output ensemble for 50 ms, 30 ms after presentation of the probe. A positive result indicated a clockwise response and a negative result indicated a counter clockwise response. As can be seen in Figure 3.4 the proportion of clockwise responses over orientation difference of our model follows a similar trajectory as those of 30 human participants who each performed 1344 trials, although it does have a slightly higher accuracy at the highest orientation differences.

4. Discussion

We developed an end-to-end spiking neural model able to execute a delayed-response task. Maintenance of information in the working memory network was realised by implementing the short-term synaptic plasticity based on calcium kinetics as proposed by Mongillo and colleagues (2008). We demonstrated that our model adheres to recent theories regarding activity-silent storage, where storage is realised by synaptic facilitation and not persistent firing. Furthermore our model harmonizes with the ‘pinging’ approach, where when a network is probed with a task-relevant stimulus, it ‘rings’ back with what is maintained in the activity-silent states. Finally we showed that our model does not just implement activity silent-storage, but that maintenance in facilitated synapses can lead to behaviour similar to that observed in humans.

Construction of the model relied heavily on the calcium-mediated STSP as proposed by Mongillo and colleagues (2008). What we aim to add to their proposal is the demonstration that a model consisting of biological spiking neurons implementing this leads not just to robust maintenance, but also leads to function. Where Mongillo and colleagues (2008) stored arbitrary items, our model maintains ‘real world’ input in the form of the experimental stimuli. Furthermore our model is able to use the maintained information to produce task-relevant responses.

4.1. Model justification

There are a number of design choices that have up till now not been justified. An obvious critique of our model might be that neurons in our model, as opposed to biological neurons, have no baseline firing rate; as is evident by the lack of any activity during the delay-period of a trial. This is a currently consequence of how the tuning curves of neurons were determined. Their intercepts were set so that neurons in the sensory and memory ensembles only fire when presented with input in order to demonstrate activity-silent storage and avoid interference with the implemented STSP. Care could be taken to implement some background firing in our model, increasing the plausibility of the model. This should however not considerably influence the functioning of our model as the maintenance of information is realised in facilitated synapses; there will be storage in activity-silent states either way.

Some additional parameter settings warrant discussion. Do the amount of neurons per ensemble and the amount of dimensions chosen matter? In general adding more neurons will improve the representation of vectors and the approximation of the functions computed over those. Increasing the number of dimensions of an ensemble expands what it might possibly be represent. Changing the number of neurons and dimensions will change the quality of the representation and will thus influence the amount of errors made during the task. However it will not change general function of the model. Other biologically relevant parameters were left to default, which are consistent with neocortical pyramidal cells (Bekolay et al., 2014). Parameters
U, $\tau_{D}$ and $\tau_{F}$ were set the same as Mongillo and colleagues (2008). $\tau_{F} \gg \tau_{D}$ and $\tau_{F}$ on the order of 1s are consistent with patch-clamp recordings of facilitating excitatory connections in the ferret prefrontal cortex (PFC; Wang et al., 2006).

One might ask how important the use of Nengo was for the creation of the model. The vector representation in Nengo provides us with an intuitive method to link our models spiking activity to representation and function. In Nengo connections are defined on the representational (vector) level, however a weight matrix of ‘end-to-end’ connections can easily be derived should one wish to create a ‘classic’ spiking network with the same behaviour as a Nengo model. Thus Nengo isn’t a prerequisite for our model, but rather, Nengo provided us with a convenient toolset to create a functioning model.

A matter that does need to be addressed is how our model relates to areas in the brain. Wolff and colleagues (2017) recorded electrical activity in the posterior region. However WM is traditionally attributed to the prefrontal cortex (PFC; eg., Fuster, 2015). The maintenance of WM related information without persistent activity has been found in both the posterior (Myers et al., 2015; Sreenivasan, Vytlacil, & D’Esposito, 2014; Wolff et al., 2017) and frontal (Barak et al., 2010; Stokes et al., 2013) region. It appears that especially sensory working memory, the short-term storage of sensory stimuli to guide behaviour, should be attributed to the relevant sensory systems (Pasternak & Greenlee, 2005; Sreenivasan et al., 2014). While the exact function of the different regions implicated in WM might differ, the neural substrate of information maintenance might be similar. In order to match the data we had, our model was designed as a model of sensory WM. However in principle it can be used as a basis for a model of the PFC; any type of information can be maintained in our model, as long as it can be represent as a vector.

4.2. Comparison to earlier models

Myers and colleagues (2015) described a neural population model consisting of a three layer architecture: a stimulus layer, template layer and a decision layer not unlike the sensory, memory and decision ensembles in our model. A big difference between the two models is that their decision layer only receives input from the template layer, while ours receives input from both the sensory and memory ensembles. The template layer in the model by Myers and colleagues (2015) acts like a match-filter, it is able to maintain a stimulus orientation, and when presented with a subsequent stimulus orientation convey the signed difference between the two to the decision layer.

Although their model didn’t consist of spiking neurons, our model might provide means to create a spiking variant of their proposed model architecture. The memory ensemble in our model can be viewed as a match-filter. After onset of the probe stimulus the orientation represented by our memory ensemble shifts to the orientation of the probe stimulus, from the direction of the orientation of memory item. This shift in represented orientation by itself indicates a degree of difference between the two orientations, including the sign of this difference. A spiking correlate to their decision layer could consist of an population of neurons that measures this shift in representation. This might be realised by an ensemble that computes a time derivative with respect to the orientation (Tripp & Eliasmith, 2010). However exploratory analysis indicated this to be less robust than our current method.

Another model with a comparable architecture to ours was proposed by Barak and colleagues (2010). Their model consisted of a sensory and memory population. After presentation of a stimulus, connections from the sensory population to the memory population will be facilitated. Subsequently, during the delay period, an increasing uniform current is applied to the network which activates the neurons in the memory population that have facilitated incoming connections. During presentation of a subsequent probe, mutual inhibition between the sensory and memory population will guide decision making. The model by Barak and colleagues (2010) is able to explain the observed ‘ramping’ up of activity during anticipation of a probe stimulus. Their model however does not consist of spiking neurons, which might provide additional complexities. Furthermore it is not clear if the gradually increasing external current is essential to extract the information maintained in facilitated connections into the memory.
representation. It can be expected that in the brain bottom-up stimulus driven activity might also be able to activate the information stored in weights, for instance when the timing of the probe is unknown.

4.3. Future directions

Mongillo and colleagues (2008) observed a bi-stable regime in their model: with added background noise, neurons with facilitated connections reactivate spontaneously. Subsequently, due to the dynamics of $u$ and $x$, the reactivated neurons will be shortly depressed before being facilitated again, leading again to reactivation. In this regime the time between subsequent reactivation is on the scale of $\tau_D$, as it is controlled by the recovery from the synaptic depression. A brief exploratory analysis shows that such a bi-stable regime also exists in our model, as can be seen in the spike raster plot in Figure 4.1. This provides our model with another method of maintain information, possibly over a longer period of time. Utilizing this method was not a requirement for our current study, however future studies, modelling more complex tasks, could employ it.

The absence of any activity during the delay period of the task in our current model directly contradicts Wolff and colleagues (2017): they were able to decode the maintained stimulus for some time after its offset. The observed activity in Figure 4.1 in might provide a possible explanation, where non-specific background input after presentation of a stimulus might temporarily push the network in this regime. Similarly the ‘ramping up’ of activity before the probe as observed by Barak and colleagues (2010) might then be explained by an increase of background input when anticipating the probe. There has recently been increasing evidence for the dynamic coding framework, which states that information maintained in the WM network traverses a highly dynamic path through neural activation (Stokes et al., 2013). It is not yet clear how this relates to our model. A possible clue might come from the model by Singh & Eliasmith (2006). Neurons in their model represent two dimension, where one dimension represents time and the other a stimulus. Their model elegantly captures a wide variety of response of observed neural responses during a WM task; the inclusion of time as a dimension represented by the neurons in the network naturally leads to a dynamic firing pattern over time.

A possible role for delay activity in WM might be tracking the focus of attention (Lewis-Peacock, Drysdale, Oberauer, & Postle, 2012). Wolff and colleagues (2017) performed a second experiment where again two memory items were presented at the start of the trial, however in this version both were eventually probed. The prioritisation of attention was guided by manipulating the order in which the items were probed. Wolff and colleagues (2017) were able to only decode the stimulus that was currently prioritised from the neural activity of the participants. This could be simulated with our model: a prioritised stimulus might be pushed into a persistent firing state as in Figure 4.1 by top-down non-specific input, while a non-prioritised stimulus might be completely maintained in activity-silent states as in Figure 3.1.

To conclude, our model shows that maintenance by means of calcium-mediated STSP can lead to function. It is broadly consistent with current theories regarding activity-silent storages in WM and is able to show a variety of effects observed during WM tasks. Furthermore it provides a solid basis to explore a model that incorporates the focus of attention by integrating both activity-silent storage and storage by persistent firing.

References


Barak, O., & Tsodyks, M. (2014). Working
Peirce, J., Gray, J. R., Simpson, S., MacAskill, M.,


