ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

MASTER THESIS:
COMPUTATIONAL MODEL OF MOUSE TEXTURE DISCRIMINATION TASK EXPERIMENT

Master Thesis
Valery Metry

Referee:

Prof. Fritjof Helmchen
Brain Research Institute, University of Zürich

Co-Referee and Supervisor:

Prof. Daniel Kiper and Dr. Yulia Sandamirskaya
Institute of Neuroinformatics, University of Zürich and ETH Zürich

End Date: September 17, 2018

Contents

1 Introduction

1.1 Computational modelling

Everything we do, think, decide and how we perceive the world is controlled and coordinated by our brain. It is the master coordinator of our lives. Although we are in the 21$^{\text{th}}$ century, the brain is still an overall mystery that is waiting to be fully understood. But how would one apprehend such a complex structure like the brain?

An approach to this is building a model. A model is per definition a simplified representation of a system over some time period or spatial extent intended to promote the understanding of the real system. There is a branch in neuroscience called computational neuroscience, which employs mathematical models, theoretical analysis and abstractions of the brain in order to gain further understanding about the principles that govern structure, physiology, development and cognitive abilities of the nervous system (Trappenberg, 2009; Sejnowski, Koch, & Churchland, 1988; Dayan & Abbott, 2001; Gerstner, Kistler, Naud, & Paninski, 2014). Computational neuroscience aims to describe biologically plausible physiology and dynamics of neurons and neural systems.

The field is relatively young, since it has emerged in the 20$^{\text{th}}$ century. Early historical roots can be traced to famous works such as those of of Lapicque, Hodgkin and Huxley, Hubel and Wiesel and Marr. 1907 Lapicque introduced the integrate and fire model of the neuron, which is a biological spiking model. The model provided a mathematical description of the properties of cells in the nervous system, that generate sharp electrical potentials across their cell membrane. This model will be further introduced in the methods section. About 40 year later, Hodgkin and Huxley developed the voltage clamp, which is an experimental method to measure ion currents through membranes of neurons, while holding the membrane voltage at a set level (Hodgkin, Huxley, & Katz, 1952). This technique lead to the creation of the first biophysical model of the action potential. Another important discovery in the field was made by Hubel and Wiesel. They discovered neurons in the primary visual cortex, which is the first area to process information coming from the retina and it has oriented receptive fields and is organized in columns (Hubel & Wiesel, 1962). A few year later Marr focused on the interactions between neurons and suggested computational approaches to study functional groups of neurons within the hippocampus and neocortex to gain further understanding on how they interact, store, process and transmit information (Marr, 1969). Rall started computational modelling of biophysical realistic neurons and dendrites with the first multicomponent model using cable theory, which is composed of mathematical models to calculate the electric current and accompanying voltage along neurites, more in particular the dendrites that receive synaptic inputs at different times and sites (Rall, 1964).

Computational neuroscientists collaborate closely with experimentalists to analyse novel data and synthesize new models of biological phenomena. The models cover a variety of topics such as single neuron modelling, developmental, axonal patterning and guidance models, sensory processing models, memory and synaptic plasticity models, behavioural models and models of cognition, discrimination and learning.

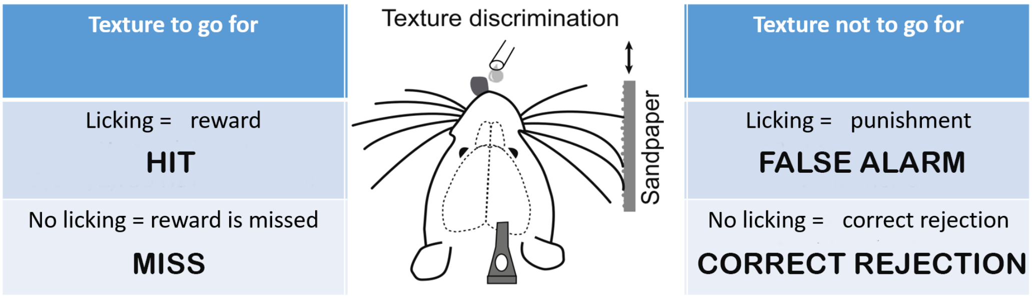
There are also different approaches to create the models, such as for instance the use of probabilistic math and different coding schemes. Examples of different coding schemes are rate coding, temporal coding, population coding and sparse coding.

This thesis focuses on modelling behavioural observation made during the texture discrimination task conducted with mice and to use an actor-critic reinforcement learning architecture in which the TD error is based on the firing activity of certain neurons in the model. In the first part of the introduction the texture discrimination task will be introduced, followed by a brief overview of whisker sensory system that is heavily involved in the texture discrimination task and the orbitofrontal cortex that is likely to be involved. The final chapter of the introduction quickly introduces reinforcement learning and the TD error.

1.2 Texture discrimination task

The texture discrimination task in rodents is popular among researchers in biology (Chen, Carta, Soldado-Magraner, Schneider, & Helmchen, 2013; Moore, 2004; Mehta & Kleinfeld, 2004; Arabzadeh, Petersen, & Diamond, 2003; Arabzadeh, Panzeri, & Diamond, 2004) and robotics (Fend, Yokoi, & Pfeifer, 2003; Seth, McKinstry, Edelman, & Krichmar, 2004; Wijaya & Russell, 2002; Kim & Moeller, 2004) as well. It serves as a framework to unravel the mystery of information coding and a vast array of literature has been generated.

The texture discrimination task is a based on operant conditioning, which is a learning process through which behaviour is modified by reward and punishment. The animals is trained to associate a particular texture with reward delivery and to suppress licking for reward when a non-target texture is presented. The omission of licking when the wrong texture is presented is often enforced by mild punishment with an unpleasant loud sound noise, time outs or delayed trial continuation (Feldmeyer et al., 2013; Helmchen, Gilad, & Chen, 2018). The experiment relies on the go/no-go paradigm as the animal is simply conditioned to go (lick) after receiving a certain stimulus (texture) and refrain from licking (no-go) for reward when a non-target distractor (different texture) stimulus is presented. The animal first needs to discern if the correct stimulus was presented and subsequently needs to make an active decision between two choices (lick or not lick). The four possible outcomes of the texture discrimination task can be seen in Figure 1.



Texture to go for	Texture not to go for
Licking = reward HIT	Licking = punishment FALSE ALARM
No licking = reward is missed MISS	No licking = correct rejection CORRECT REJECTION

Figure 1: Go/ No-go texture discrimination task in a nutshell. Mouse is headfixed and receives the roughness of a certain texture as sensory input and subsequently needs to take a decisions whether to lick or not for reward. The four possible outcomes are depicted on the figure. The mouse illustration is taken from (Helmchen et al., 2018).

1.3 Whisker sensory system

Rodents such as mice and rats use their whiskers (vibrissae) to gather information about their surroundings. As they are nocturnal animals, they mostly live in environments with dim or fading light and as a consequence they need to rely on their tactile information. The whisker sensory system is likely to have evolved to compensate for the poverty of visual information. Rodents have well-developed a number of tactile capacities. By the use of their whiskers they are able to obtain information about object features such as size and shape (Brecht, Preilowski, & Merzenich, 1997), object location (O'Connor et al., 2010), object texture (Carvell & Simons, 1990; Von Heimendahl, Itskov, Arabzadeh, & Diamond, 2007; Carvell & Simons, 1990; Guić-Robles, Valdivieso, & Guajardo, 1989; Prigg, Goldreich, Carvell, & Simons, 2002), object distance (Shuler, Krupa, & Nicolelis, 2001; Solomon & Hartmann, 2006), bilateral distance (Knutsen, Pietr, & Ahissar, 2006; Krupa, Matell, Brisben, Oliveira, & Nicolelis, 2001) and orientation (Polley, Rickert, & Frostig, 2005). Tactile exploration consists of an interplay of motor output and sensory input as the rodent uses a sweeping motion of the whiskers in a rhythmic forward-backward cycle (Berg & Kleinfeld, 2003; Hill, Bermejo, Zeigler, & Kleinfeld, 2008). The whiskers are moved in rhythmical sweeps at frequencies ranging between 5 Hz and 15 Hz (Carvell & Simons, 1990; Kleinfeld, Sachdev, Merchant, Jarvis, & Ebner, 2002). Due to the rhythmical nature of the whisking neuronal activity in the primary somatosensory (barrel) cortex varies rhythmically in synchrony with whisker movement and contact against obstacles (Mégevand et al., 2009; Fee, Mitra, & Kleinfeld, 1997; Crochet & Petersen, 2006; Von Heimendahl et al., 2007).

The whiskers are made of inert material, tactile sensitivity is provided by an array of follicles at the mystical pad that anchor the whiskers to the skin (Dörfl, 1985; Diamond, Von Heimendahl, Knutsen, Kleinfeld, & Ahissar, 2008) and each follicle is innervated by peripheral branches composed of about 200 cells of the trigeminal ganglion (Dörfl, 1985). The nerve endings of the trigeminal ganglion convert the mechanical energy into action potentials that travel past the cell bodies in the trigeminal ganglion and continue along the central branch to excitatory glutamatergic synapses in the trigeminal nuclei of the brainstem (Torvik, 1956; Clarke & Bowsher, 1962). Afferent vibrissal information is then conveyed to the thalamus, where trigeminothalamic neurons of the principal trigeminal nucleus are arranged in a barrel like divisions which are called barrelets. Each barrelet receives strong input from a single whisker (Veinante & Deschênes, 1999). The principal trigeminal neurons then project to the ventral posterior medial (VPM) nucleus of the thalamus, which also contains barrel like structures called barreloids. VPM neurons respond precisely and rapidly to whisker deflection with the particularity that one whisker evokes stronger response then the others (Friedberg, Lee, & Ebner, 1999; Simons & Carvell, 1989; Brecht & Sakmann, 2002). The axons from the VPM neurons within individual barreloids lead to the layer 4 barrel field of the primary somatosensory cortex (Deschênes, Timofeeva, Lavallée, & Dufresne, 2005). In this layer 4 barrels field each whisker is represented by a discrete and well-defined structure (Woolsey & der Loos Van, 1970) and interestingly these layer 4 barrels are somatotopically arranged in an almost identical fashion to the layout of the whiskers on the snout (Petersen, 2007).
Motor movement is controlled by the vibrissal motor cortex (Brecht, Schneider, Sakmann, & Margrie, 2004) and is interconnected with cortical and subcortical sensory structures (Miyashita, Keller, & Asanuma, 1994) and the activity in the motor cortex is also modulated by whisker movement (Kleinfeld et al., 2002; Chakrabarti, Zhang, & Alloway, 2008).

Knowledge on this sensory system permits to predict behavior as it can be generated by signals originating from whisking. For instance spike counts in the barrel cortex neurons, integrated over seconds of time resulted in a good prediction of rat performance when they perceived pulsative whisker stimuli. (Gerdjikov, Bergner, & Schwarz, 2017). Due to the neuroethological relevance, the rodent whisker-system is a popular model system for studying tactile information processing (Feldmeyer et al., 2013; Petersen, 2007). Experiments with head-restrained rodents enable precise tracking of behavioural parameters resulting from whisking as the animal is not able to move anything else. These experiments are used for precise tracking of behaviour parameters such as whisker touch and movement and generate data about the neuronal activity from the cellular to the large network level by intercellular recordings, calcium imaging techniques and optogenetics (Helmchen et al., 2018). Diverse whisker-based discrimination tasks experiments have been created for head-restrained animals (Guo et al., 2014). In the texture discrimination task the mechanism by which the animal can distinguish between different textures are so-called stick-slip events. Sandpaper for instance has strong textures and the whiskers get caught by the sandpaper grains, get streched and then released like a spring (Arabzadeh, Zorzin, & Diamond, 2005; Wolfe et al., 2008; Boubenec, Shulz, & Debrégeas, 2012). The frequency of the slip-stick events encodes for graininess. It is beneficial for the animal to engage in active whisking as it increases the likelihood of stick-slip events in a texture dependent manner (Von Heimendahl et al., 2007; Zuo, Perkon, & Diamond, 2011; Chen et al., 2015). Read-out parameters in this experiment are whisking angle, that can be measured for individual whiskers or aRveraged across multiple whiskers, whisker set point, contact-induced curvature change and frequency of stick-slip events. These parameters allow to estimate the lateral and axial forces that act on the whisker follicles (Von Heimendahl et al., 2007; Wolfe et al., 2008; Chen et al., 2015; Boubenec et al., 2012; O'Connor et al., 2010; Pammer et al., 2013). The representation of touch events in the neuronal population of the S1 barrel cortex are investigated using electrophysiological recordings or calcium imaging of touch evoked neuronal responses. Imaging during discrimination task revealed coordinated patterns of activity between S1 and S2 neuronal populations, to both motor behavior (whiskering and licking) as well as sensory processing (Chen, Voigt, Javadzadeh, Krueppel, & Helmchen, 2016). Further imaging results in anaesthetized rats hint that whisker vibrations associated with different textures evoke cortical responses that differ according to the texture. Grainier textures evokes more spikes per sweep (Arabzadeh et al., 2005; Arabzadeh, Panzeri, & Diamond, 2006).

1.4 Orbitofrontal cortex and decision making

Another brain region which might be involved in the texture discrimination task is the orbitofrontal cortex (OFC), because the ability to maintain information to be manipulated and integrated with other information to guide behavior has been described as working or representational memory and depends on the OFC (Goldman-Rakic, 2011). The OFC is interconnected with limbic areas, those areas are known to support a variety of functions including emotion, motivation, long-term memory, behavior and olfaction. This interconnection permits the OFC to allow information regarding outcomes to access representational memory (Schoenbaum & Roesch, 2005). Studies in humans have shown that a neural correlate of expected outcome value is present in the OFC, because the blood flow in the OFC changes during anticipation of outcomes and when the value of the outcome is modified (O'Doherty, Deichmann, Critchley, & Dolan, 2002; Gottfried, O'doherty, & Dolan, 2003). Further evidence in primates and rodents suggest that neural activity in OFC preceding predicted rewards and punishments reflect the value of those outcomes (Schoenbaum, Setlow, Saddoris, & Gallagher, 2003; Schoenbaum, Chiba, & Gallagher, 1998). The OFC has a critical role in signalling outcome expectancies in reinforcer devaluation tasks. These are tasks that assess the control of behaviour by an internal representation of the value of an expected outcome. Assume that rats have been trained to associate a particular texture with reward, when this texture however yields to punishment all of the sudden, then it is devalued and when the texture is presented again in another trial the rat will respond less to the cue than the non devalued cues. This decrease in responding happens in addition to normal decrease caused by extinction, leading to faster unlearning. Experiments have shown that OFC lesioned rats fail to change their behaviour after devaluation. First rats were trained to associate a light cue with food and later the food was devalued by pairing it with an illness. OFC lesioned rats condition and devalue normally, but unlike OFC intact rats do not show the effect of devaluation on conditioned responding (Gallagher, McMahan, & Schoenbaum, 1999). They continue to respond to light to obtain food, even though they will not eat it when presented. This effect can be observed whether OFC lesions are made before or after learning, this indicates that OFC is not solely involved in acquiring the cue-outcomes association (Pickens et al., 2003). Rather it is thought that the OFC is critical to control conditioned responding according to internal representations of the new value of the expected outcome (Izquierdo, Suda, & Murray, 2004). Further experiments with rodents in which genes in the OFC were selectively silenced or lesioned also showed impaired reversal learning (Ferry, Lu, & Price, 2000; Kolb, Nonneman, & Singh, 1974) and (Banerjee et al. 2018, unpublished data). As mentioned above OFC lesions do not disrupt the ability to perform discrimination but there is a deficit in the early stage of reversal learning, which is characterized by the inability to inhibit previous reinforced response, which lead to preservation. A finding that has been found in primates (Rolls, Hornak, Wade, & McGrath, 1994; Dias, Robbins, & Roberts, 1996; Iversen & Mishkin, 1970; Jones & Mishkin, 1972) and rats (Brown & Bowman, 2002; Chudasama & Robbins, 2003; McAlonan & Brown, 2003). The deficit of the OFC lesioned organisms might be due to failure of prepotent instrumental response. This failure in inhibition might be enhanced by effects of proactive interference from previous established association, leading to an enhanced expression of stimulus response habit that tends to be impervious to changes in value of reinforcement (Balleine & Dickinson, 1998; Dickinson & Balleine, 1994; Boulougouris, Dalley, & Robbins, 2007). Proactive interference is when the old memory information prevents the recall of newer information.

1.5 Reinforcement learning

The computational field of reinforcement learning has provided a normative framework within which decision making can be analysed and conditioned behavior can be understood (Sutton & Barto, 1998). Reinforcement learning emphasises on the question on how to map situations to actions to maximize the reward signal and to minimize punishment. The learner is not told what actions to take but must discover which actions yield reward by trying them, in order to do so the agent must be able to sense the state of the environment and must be able to take an action that affects the state. Optimal action selection is based on predictions of long term future consequences, such that decision making is aimed at maximizing rewards. Neuroscientific evidence provided by various animal experiments such as lesion studies, pharmacological manipulations and electro-physiological recordings have provided links to neural structures similar to key computational constructs in reinforcement learning as for instance the neuromodulator dopamine. Dopamine provides basal ganglia target structures with phasic signals that convey a reward prediction error, which influences learning and action selection in stimulus driven behavior (Schultz, 1998; Houk, Davis, & Beiser, 1995; Houk et al., 1995). Operant conditioning such as for instance the texture discrimination task introduced in the general introduction, involve learning to select actions that will increase the probability of rewarding events and decrease the probability of aversive events (Thorndike et al., 1912; Skinner, 1935). From a computational point of view, such decision making can be treated to optimize the consequences of actions in terms of some long-term measure of total obtained reward and avoided punishment and thus it makes the study of operant conditioning represent an inquiry into the most fundamental form of decision making (Niv, 2009). The texture discrimination task can be addressed as a reinforcement problem. This task has been previously described and can be summarized as follows. A particular texture (environmental state) is presented to the animal and the animal needs to make a decision based on its sensory information and take an action (that affects the state of the animal), which is lick or not to lick in order to maximize reward and minimize punishments.

1.6 Aim of the thesis

The aim of this thesis is to create a computational model of the texture discrimination task using Brian2, which is a spiking neural network simulator that will be further introduced in the methods section. The key idea of this project is to link the ideas of reinforcement learning with spiking neural networks to create a model that can simulate learning in mice and investigate on how a particular stimulus is mapped to an outcome and learned with an actor critic implementation.

2 Methods

2.1 Spiking neural networks simulator Brain2

Computational simulations have become an important tool in neuroscience and Brian2 is used to build the model. It is a free, open source simulator for spiking neuronal networks that can be used across multiple platforms. It is written in Python and focuses on simplicity, extensibility of neuronal and synaptic models. The models can be described using mathematical formulas with the use of physical units (Goodman, Stimberg, Yger, & Brette, 2014; Stimberg, Goodman, Benichoux, & Brette, 2014). Brian2 is valuable for working on non-standard neuronal models, that cannot be easily covered by other existing software and it is an alternative to Matlab or C simulations. It has a simple syntax and is also well suited for starting into computational neuroscience (Goodman & Brette, 2008). Simulating a neural model means tracking the change of neural variables such as membrane potential and synaptic weights over time. The rules governing these changes take two principal forms: one of continuous updates as for example of the decay of the membrane potential back to a resting state in absence of inputs and the second of event-based updates as for instance the reset after a spike in an integrate and fire neuron. An event can be described as a change in the state variables of the system that is triggered by a logical condition on these variables. Continuous updates are described by deterministic or stochastic differential equations while event-based updates are described as a series of mathematical operations (Stimberg et al., 2014). The description of an accurate model needs both variables that evolve continuously and discontinuously through events. This can be achieved by the use of the leaky integrate and fire model (LIF).

2.2 The leaky integrate and fire neuron model (LIF)

In this thesis the LIF with noise was used.
The most simple version of LIF can be described as follows:

$$dv/dt = -(v - v_0)/\tau_m$$

$$\text{After} \quad v > v_{th} : v \leftarrow v_0$$

where v_0 is the resting and reset potential of the cell and τ_m is the membrane time constant. If the voltage v is bigger than the threshold v_{th} than v will be reset to v_0. At the moment the threshold is reached a spike is emitted.
LIF has different states as for instance excitable and non excitable (refractory) state. An event (spike) can trigger changes in the state variable and also a transition between these states. The fact that a neuron is not able to generate a second action potential for a short time after the first one is emitted, is modelled by imposing a refractory period after each spikes.
In order to include a random element to the equation to simulate noise the symbol ξ was added to the differential equation of the LIF:

$$dv/dt = -(v - v_0)/\tau_m + \xi * \tau^{-0.5}$$

The symbol ξ stands for a stochastic differential and behaves similar to a Gaussian random variable with mean 0 and standard deviation 1. It is also taken into account how stochastic differentials tend to scale with time, that is why ξ is multiplied by $\tau^{-0.5}$. The integration method used for the differential neuron equation in the simulation is the Euler-Maruyama method. For further information about stochastics and the Euler-Maruyama method the reader is referred to the following textbook (Mao, 2007).

2.3 Synapses and learning rules

Synapses connect a presynaptic neuron to a postsynaptic neuron and events (spikes) can trigger changes in pre- and postsynaptic neural variables. In contrast to neural models, there is no need for a synaptic threshold condition since action potentials are emitted from the pre- and postsynaptic neurons according to the threshold conditions of the neural model. As spikes occur in the presynaptic neuron, it causes an instantaneous change.

$$v_{post} \leftarrow v_{post} + w$$

In simple synapses a variable w is added to the voltage of the postsynaptic neuron v_{post} when spikes occur in the presynaptic neuron.
More complex synapses are governed by another learning rule, in our case the Fusi learning rule. The Fusi learning rule is an extension of the model of spike time dependent plasticity (STDP). Thus STDP will be introduced first.

$$\Delta w = \sum_{t_{pre}} \sum_{t_{post}} w(t_{post} - t_{pre})$$

The change of the synaptic weight w is the sum over all presynaptic spike times t_{pre} and postsynaptic spike times t_{post} of some function W of the difference in these spike times (Gerstner, Kempter, van Hemmen, & Wagner, 1996; Kempter, Gerstner, & Van Hemmen, 1999). $W(\Delta t)$ stands for the so called learning window and can be defined as follows:

$$W(\Delta t) = A_{pre} e^{-\Delta t/\tau_{pre}} \quad \Delta t > 0$$
$$W(\Delta t) = A_{post} - e^{\Delta t/\tau_{pre}} \quad \Delta t < 0$$

The parameters A_{pre} and A_{post} depend on the current value of the synaptic weights and the time constant is in the order of $\tau_{pre} = \tau_{post}$.
In other words when a presynaptic spike shortly precedes a postsynaptic action potential, it is likely that depolarization of an LIF neuron is high, resulting in long term potentiation (LTP) and memory consolidation. If the presynaptic spikes comes shortly after the postsynaptic action potential, the postsynaptic neuron is likely to be hyperpolarized, which results in long term depression (LTD).

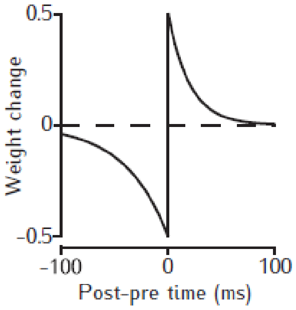


Figure 2: Illustration of the STDP kernel.

The Fusi learning rule differs slightly from STDP. It has another variable which is the calcium variable $C(t)$ with a long time constant that is a function of postsynaptic spiking activity (Brader, Senn, & Fusi, 2007).

$$\frac{dC(t)}{dt} = -\frac{1}{\tau_C} + Jc\sum_i \delta(t - ti)$$

where the sum over postsynaptic spikes arriving at times ti. Jc is the contribution of a single postsynaptic spike and τ_C is the time constant. The variable $X(t)$, which has a similar function to the A parameters of the STDP, is modified on the basis of the postsynaptic depolarization v(t) and the postsynaptic calcium variable C(t). The synapses are bistable with efficacies J_+ (potentiated) and J_- (depressed). The efficiencies J_+ and J_- can be any number and are not restricted to binary values (0,1). The internal state of the synapse is represented by X(t), and the efficacy of the synapse is determined whether $X(t)$ lies above or below a threshold θ_x. The variable X(t) is restricted to the interval $0 \leq x \leq X_{max}$ and is a function of C(t) and of both pre- and postsynaptic activity. A presynaptic spike which arrives at t_{pre} reads the instantaneous values of V_{pre} and $C(t_{pre})$ and the conditions for a change in X depend on those values as follows

$$X \to X + a \quad if \quad V(t_{pre}) > \theta_v \quad and \quad \theta^l_{up} < C(t_{pre}) < \theta^h_{up}$$

$$X \to X - b \quad if \quad V(t_{pre}) \leq \theta_v \quad and \quad \theta^l_{down} < C(t_{pre}) < \theta^h_{down}$$

where a and b represent jump sizes, θ_v is the voltage threshold and θ^h_{up}, $theta^l_{up}$, $theta^h_{down}$ $theta^l_{down}$ are the thresholds on the calcium variable. In absence of a presynaptic spike or if the previously above mentioned conditions are not met, X(t) drifts toward one of the two stable values,

$$\frac{dX}{dt} = \alpha \quad if \quad X > \theta_x$$
$$\frac{dX}{dt} = -\beta \quad if \quad X \leq \theta_x$$

where α and β are positive constants and θ_X is a threshold on the internal variable. If at any point during the time course $X < 0$ or $X > 1$, then X is held at the respective boundary value. The efficiency of the synapse is determined by the value of the of the internal variable at t_pre. If $X(t_{pre} > \theta_X)$, the synapse is potentiating and in the case that $X(t_{pre} \leq \theta_X)$, the synapse is depressing.

2.4 Temporal difference (TD) learning

The framework of reinforcement learning provides a theory and algorithms for learning (Sutton & Barto, 1998). An attractive formulation of reinforcement learning is temporal difference TD learning (Sutton, 1988). TD learning assumes that an agent moves between states in its environment by choosing appropriate actions in discrete time steps. Rewards are given in certain conjunctions of states and actions, and the agents goal is to choose its actions in order to maximize the amount of reward it receives (Frémaux, Sprekeler, & Gerstner, 2013). In TD learning, the goal of the learning system (also referred to as the agent), is to estimate and predict values of different situations or states, in term of future reward or punishments (Niv, 2009). From a learning standpoint, the TD model assumes that the goal of a mouse in the texture discrimination task is to learn the value of the texture (stimulus) that will lead to reward after licking at the water delivery port. One way to do this is to estimate for each texture the amount of reward that the mouse can expect to receive in the future. To further quickly introduce TD learning think of a Markov chain. A Markov chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. In this Markov chain different states follow one another in a predefined probability distribution $P(s_{t+1}|s_t)$. A useful quantity to predict in such a situation is the expected sum of all future rewards, given the current state S_t, which can be referred as the value of state S_t.

$$v(S_t) = E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ...|S_t] = E[\sum_{i=t}^{\infty} \gamma^{i-t} r_i]$$

$\gamma \leq 1$ discounts the effect of rewards distant in time on the value of the current state. The discount factor rate was introduced to ensure that the sum of future rewards is finite and it supports the fact that animals prefer earlier rewards to later ones. Such exponential discounting is similar to the assumption of a constant interest rate per unit time on the received rewards. To learn the values it is necessary that consistency holds only for correct values, which are those that predict the expected discounted sum of future values. If the values however are incorrect, there will be a discrepancy between the predicted values and the correct one. This is called the TD prediction error and is defined as follows

$$\delta_t = P(r|S_t) + \gamma \sum_{S_{t+1}} P(S_{t+1}|S_t)V(S_{t+1}) - V(S_t)$$

The prediction error is a natural error signal with the aim of improving the estimates of the function $V(S_t)$. The larger the error is, the larger is the difference between the expected and actual reward. When this is paired with a stimulus that accurately reflects the future reward, the error can be used to associate the stimulus with future reward.

In this project the TD error is used to gate the plastic learning synapses by scaling the learning. If the TD error is high, strong learning happens if it is low and the prediction is correct then few or no learning happens because the TD error is small. To gate learning through the TD variable, the activity of the TD error neurons was monitored with a variable that increases with neural activity and decays with time when no activity happens. The variable is activity variable. It has similar dynamics to the LIF but differs slightly by having no threshold condition and reset.

$$dA/dt = -A/\tau_A$$

The variable has been monitored in the TD error neurons and summed up to one neuron, where it is normalized such as that the maximum activity of the TD error neurons represent 1, the minimum 0 and the activity values in between minimum and maximum represent a range of scalars from 0 to 1. The Fusi equation at the plastic synapses is multiplied by this value and is thus gradually gated (Figure 2).

3 The computational model

It is crucial to introduce the structure of the model, as the structure determines what the model is doing and what the limitations are. Figure 1 illustrates the structure of the model. The first neuron group of the model is called input neurons and comprises two poisson input generator neurons.

Those neurons represent the two possible input textures, which can be presented to the mouse. As only one texture can be presented to the mouse at a given time, those two neurons are never simultaneously active. At the beginning of each trial one of the two is randomly selected to be active for the duration of the trial. The trial structure will be introduced in a later subsection of this thesis. Those two input neurons are connected to a middle layer with a Gaussian probability and as a consequence the middle layer is differentially affected by the two input neurons. As for instance the blue input neuron is stronger connected to the middle neurons in the left of the middle neuron group and thus affects them more then the neurons on the right of the neuron group and lead to higher firing in the left when the blue input is activated. Those two Gaussians thus ensure that the activity distribution across the middle layer is different when the first input (blue) is active and when the second input (orange) is active. This idea behind this difference in activity of the middle layer is that different inputs (textures) might be differentially processed by neurons. The middle neurons are connected via plastic synapses to the output neurons.

The output neurons comprise two subpopulations, which represent the two different behavioural outcomes of the texture discrimination task, that are licking and no licking. As introduced in the methods section those plastic synapses update their synaptic weight when the previous neurons group spikes shortly before the next neuron group spikes, in this case it is the middle and output group. The reason why the Gaussian distribution of the inputs of the synaptic weights do not overlap is that if they would overlap then a portion of the middle neurons would spikes regardless of which input is active, because it is connected to both inputs and thus its constant spiking would interfere in the learning of the outputs. As it always spikes it would have the tendency to associate one output with both inputs. This problem was encountered during the early stages of the implementation of the model and thus the Gaussian were implemented such as to have a minimal overlap. The synaptic weights are set to be a value between 0 and 1. When initiated their value is around 0.5, giving them the possibility to either potentiate or get depressed. These plastic synapses allow the mapping of the input to an output. The behavioural response of a mouse in the texture discrimination task is either lick or no-lick, two distinct decisions that cannot take place in the same time. To ensure that the model captures this fact a winner take all is implemented between the two subpopulations of the output group. This winner take all is the result of a synaptic connection between the two subpopulations that inhibits both subpopulations equally strong. The subpopulation that receives more input however can escape this inhibition because the excitation that is coming from the middle layer is stronger then the inhibition provided by the other subpopulations and thus that subpopulation would ”win”.

Each output subpopulation is connected to an additional neuron belonging to the binary neuron group, this neuron is used to represent the activity of the whole output neuron subpopulation that is connected to it. It can be seen like a relay station that sums up the activity of the output subpopulations. The name binary was used as designation because only one of the neurons is meant to be active for the same reasons as for the output group mentioned above. To ensure this behavior, inhibition in a winner take all fashion between the binary neurons has been implemented.

The binary neurons have a special role in reward signal generation in the programming code. They have been used to decide when the input is applied to the reward neurons. The program checks which input is active and also which binary neuron is more active. At the beginning of each simulation the rewarded connection of inputs (textures) and outputs (lick or not lick) is chosen to be rewarded. For the results showed later the connection that is to be learned is input 1 - output 1 (blue and violet in figure 3) and input 2 - output 2 (orange and green in figure 3). The input 1 to output 1 is labelled as Hit in the later sections because it represents the association of a texture with licking, while the connection from input 2 to output 2 is called as a correct rejection because it represent the omission of licking. In this model both of these connections are rewarded, because they represent correct behavioural outcomes that the mouse is required to learn during the texture discrimination task. Note that in the actual task the mouse is only rewarded when the the mouse licks for the target texture, in our case input 1 and is not rewarded when it does not lick after sensing input 2 (texture 2). For the sake of simplification the correct association of texture 2 with not licking is also rewarded, because from a conceptual standpoint it can be seen as a rather favourable situation as it has no punishment as a consequence for the mouse. Additionally it is necessary in this model to reward this situation or else it wouldn’t be learned by the plastic synapses. The learning in these synapses are gated by the TD error that is only active when reward is received. Currently the model checks what input is active and with what output it is associated. If the correct binary neuron that is connected to the output group subpopulation that harbours reward when associated with the currently active input is 2 times more active then the other binary neuron that is associated to the other output subpopulation, then the reward neurons are given an input.

While the reward is given and the reward neurons are active, they excite the TD error neurons. Those are the neurons that are used to calculate the TD error in order to gate the learning at the plastic synapses. First their activity is summed up to one neuron and subsequently this activity is normalised between 0 and 1 and this will be used to gate the learning. This will be better explained later.

Back to the binary neuron group. This group is connected to the four outcomes neurons. As the name of this neuron population suggest, those neurons represent the four possible outcomes that can be observed in the texture discrimination task. Those are Hit, Miss, Correct Rejection and False Alarm. Their spiking helps visualize which connection has been selected by the system. The way how the neurons represent that, is that they receive input from the input group and also from the binary group and the four outcomes subpopulations are also connected in a winner take all fashion, which leads to the winning of only the selected outcome. For the sake of simplicity of illustration the figure 2 only displays inhibition between neighbouring subpopulations, but in fact all the subpopulations inhibit each other.

Those subpopulations in turn are connected to a critic neuron group via plastic synapses so that the synapses can associate different synaptic weights to the critic depending on the selected outcome, that is important for the calculation of the TD error. The critic neurons inhibit the TD error neurons and is also in turn excited by the TD error neuron.

To further understand how this works it is important to examine the loop of reward neurons, TD error neurons and critic neurons. The activity of the TD error neurons is the result of a balance between excitation from reward neurons and inhibition of critic neurons. The goal of the system is to maximize the reward in a long term and to do so the correct association must be learned or in other words the plastic synapses connecting the correct associations need to be potentiated while the other synapses should depress. To do so it is important to make a prediction of the value of the current decision and update the value function accordingly. The critic represent the current value. To illustrate how this TD error works, please look at figure 3 and assume a naive system that has not been rewarded yet. In that case the synaptic weights of the plastic synapses are rather low. The input would excite the middle layer which would excite one of the output subpopulations, as those synaptic weights are low however the output group would not spike significantly, the exciting signal is then propagated to the other neurons groups until it arrives the critic. Again there are plastic synapses between the four outcomes group and the critic and those would also excite the critic minimally. Thus the critic would not spike much. In case the correct association is given, the input current of the reward neurons is switched on and the reward neurons excite the TD error greatly while the critic that would continue to spike with a similar low activity and not inhibit the TD error neurons much, which has as a consequence that the TD error neurons are very active. Due to the fact they are very active and their activity is used to gate the plastic synapses the synapses would update their synaptic weight a lot due to a high TD error . With time as the system receives more and more reward, the synaptic weights for the correct association are more and more updated, while the synaptic weights for the unrewarded association go down. This change of synaptic weights has as a consequence that the critic receives more input when reward is given and thus in turn inhibits the TD error stronger. The excitatory input of the reward neurons stays always the same and as the inhibition of the critic neurons rises, the activity of the TD error goes down, which leads to a lower value of the TD error gating and thus less learning for the plastic synapses. If the spiking and of the critic and reward neurons synchronises and the inhibitory and excitatory connections affect the TD error neurons equally strong, then the learning is off. The TD error neurons receive an additional current that gives them a baseline firing so that the TD error can be always calculated, this baseline firing is subtracted from the actual TD error calculation.

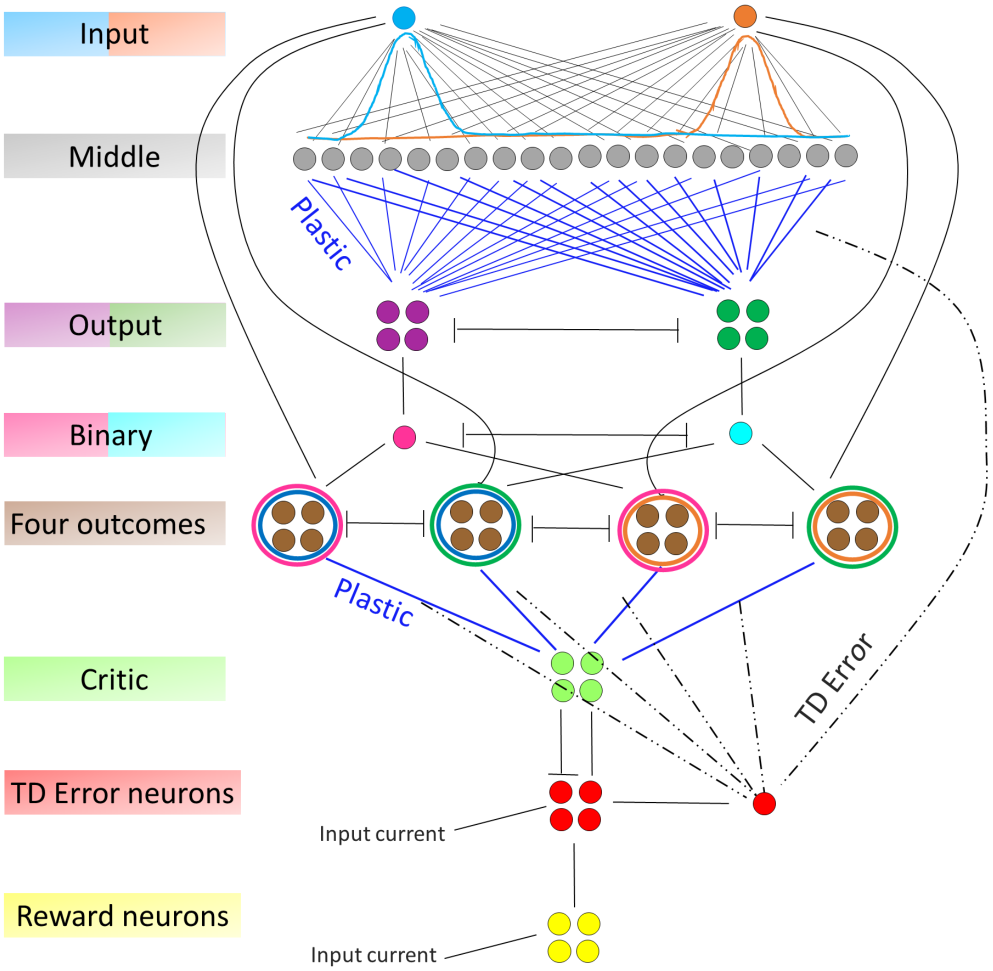


Figure 3: The architecture of the model. The model consists of 9 neuron groups, indicated in different colors. Excitatory synapses are marked with black lines, inhibitory synapses with black lines which end with vertical lines and plastic synapses which are TD error gated are marked in blue. Synaptic connection are not bidirectional and go from the upper neuron groups to the neuron groups below except the connection from the TD error neuron to the critic which excites the critic. The TD error gating is represented by black dashed lines and the connections originate at the red TD error neuron that summarizes the whole activity of the TD error neuron group and affects all the plastic synapses in the system. A winner take all is implemented between the different subpopulations of the output neurons, binary neurons and the four outcomes.

3.1 Early models

The above presented architecture is the final architecture with reinforcement learning implemented thanks to the TD error gating. The early builds of the model consisted of an unsupervised and a supervised version of the current model. This early model only contained the input neurons, middle neurons an the output neurons (Figure 3). In the unsupervised version of the model the system selects randomly an output subpopulation to be associated with the inputs. The mapping from input to output are not determined and each output is equally likely to be associated with an input. The plastic synapses were connected by the STDP and Fusi learning rule as both of them performed equally the Fusi learning rule was chosen to be tuned for the final architecture (Figure 3), because it is compatible with the Dynapse and would facilitate a possible future implementation on neuromorphic hardware. The supervised version of the model is slightly different, it has in addition to the 3 first neurons groups another neuron group that excited the output subpopulations. This neuron group always excites a given output subpopulation when a particular input of choice is active. This excitation ensures that the subpopulation that is wished to connect to a input population spikes when this input spikes and as the spike at they fire together, they wire together. That means the synaptic synapses update their weight for this association. Biologically speaking it might be unlikely that neurons learn in an unsupervised or supervised way as there is dopamine. Dopamine is neurotransmitter that signals the value of an outcome which motivates the organism to achieve an outcome (Berridge & Robinson, 1998). Considering dopamine signalling, it seems much more probable that the brain learns through a mechanism such as TD error.

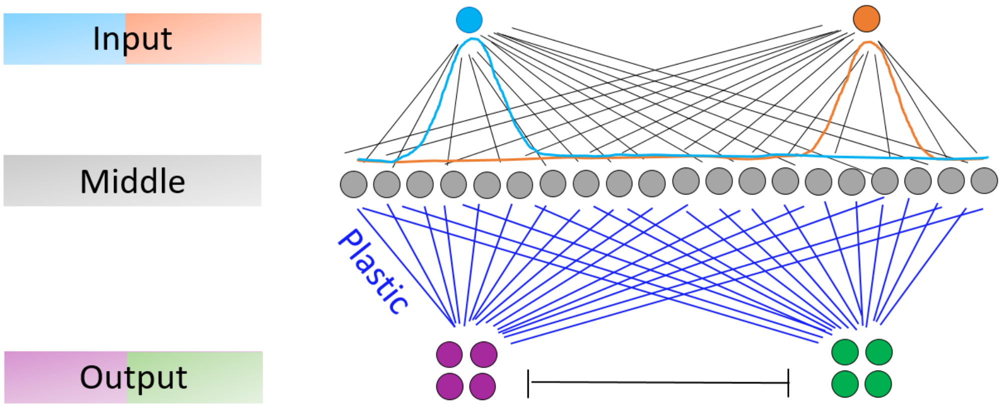


Figure 4: The architecture of an unsupervised model.

3.2 Trial setup

All trials have an equal length and are separated by a pause of equal length as the trial. At the beginning of each trial an input is randomly chosen to be active for the duration of the trial. During the pauses the inputs are turned off and as the system loses inputs all the neurons relax, which means their activity goes down. At the start of each trials the voltages and activity of all the neurons are reset to 0.

3.3 Results

3.3.1 Showcasing a simulation

In this section a simulation will be showcased. This simulation has 300 trials and 300 pauses in between trial. Each trial and pause have the duration of 1 second. At the beginning of each pause the activity of the neurons is reset. During the first 100 trials no reward is given to the system to let it explore different options and after the 100 trials they system is rewarded when input 1 - output -1 is associated (Hit) and input 2 - output 2 is associated (Correct rejection). Other associations such as input 1- output 2 (Miss) and input 2 - output 1 (False alarm) are not rewarded. Figure 5 shows the spiking pattern of all neurons during a simulation and figure 6 shows the associated legend. As this graph is fairly crowded, figure 7 shows a version that is zoomed in at the position where the transition happens between the non rewarded state and the rewarded state. The transition happens after 200 seconds. In the non rewarded phase the TD error gating is off and as reward is given to the system the TD error gating starts to affect the synapses. The black line at the top figure 7 shows the strength of the TD error gating. At times when reward is given (those are in between the light grey lines on the figure 7) the TD error gating affects the synapses the most as it is the time when the reward neurons receive a current and excite the TD error neurons.

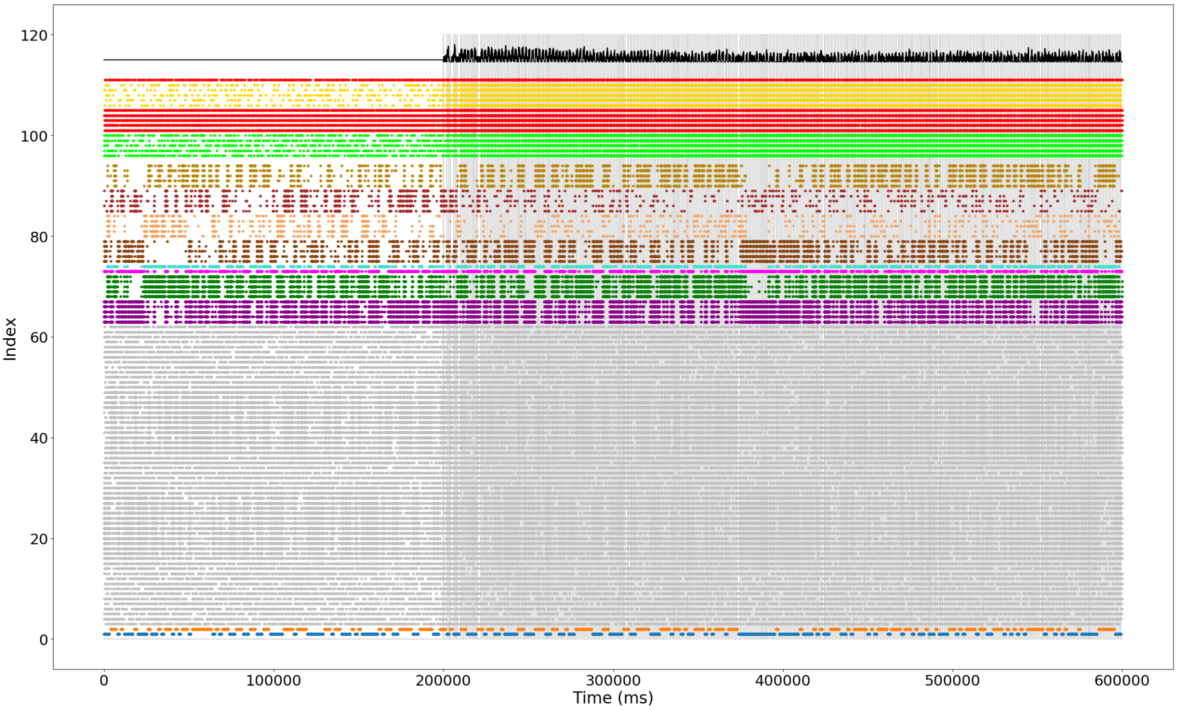


Figure 5: Spiking pattern of a simulation comprising 300 trials and 300 pauses in between. Each trial has a duration of 1s. The legend of the figure is indicated in the figure below.

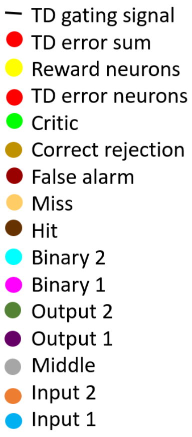


Figure 6: Legend of the figures 5 and 7. Showing the name of the neuron groups and indicting which color is associated.

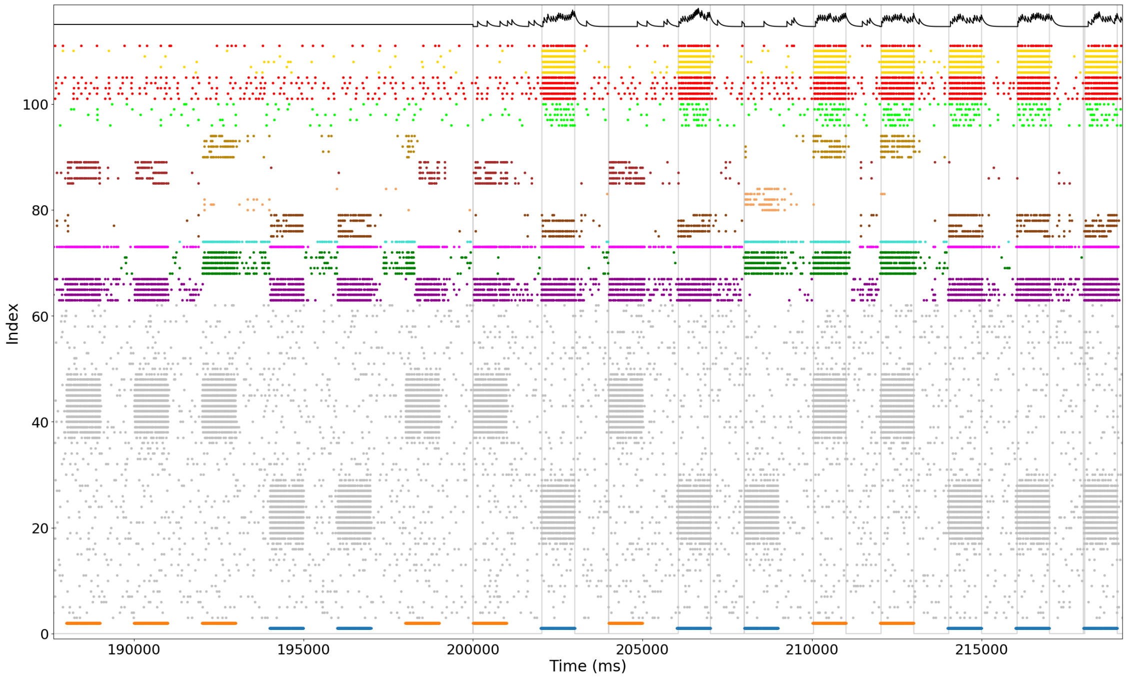


Figure 7: Spiking pattern of the transition between the non rewarded beginning and the rewarded period. The grey lines indicate when the reward is given to the system. The rewarded association is input 1 - output 1 and input 2 - output 2.

Figure 8 and 9 show the evolution of the synaptic weights of the plastic synapses during the simulation. The black line represents the border between the the beginning where no reward was administered and the moment from which on reward was given for the correct associations. On both graphs in the beginning during the naive stage (non rewarded stage), no learning happened as the synaptic weights do not rise. For the synapses between the middle and output group the synaptic weights stay relatively the same as opposed to the plastic synapses between the four outcomes neurons and the critic neurons. The reason for this slight decay of synaptic weights between the four outcomes neurons and the critic neurons is that those synapses do not get a lot of input as the critic fires relatively low. Figure 10 shows this low firing of the critic. This low firing is a consequence of low synaptic excitement to the output neurons. This is changed after the system is rewarded. During the first times the system is rewarded, the critic neurons continue to fire at a low rate while the reward neurons fire at a much higher rate when activated at the moment the correct association is rewarded. This discrepancy in firing rate has a consequence that the TD neurons are not much inhibited by the critic neurons but strongly excited by the reward neurons, thus the firing rate of the TD neurons is high. This high firing rate is used to gate the learning in the plastic synapses. As it is high the synaptic weights are strongly updated. After the beginning of reward administration the synaptic weights rise quickly (Figure 8 and 9), this is due to the high TD error firing rate shortly after the first reward administration.

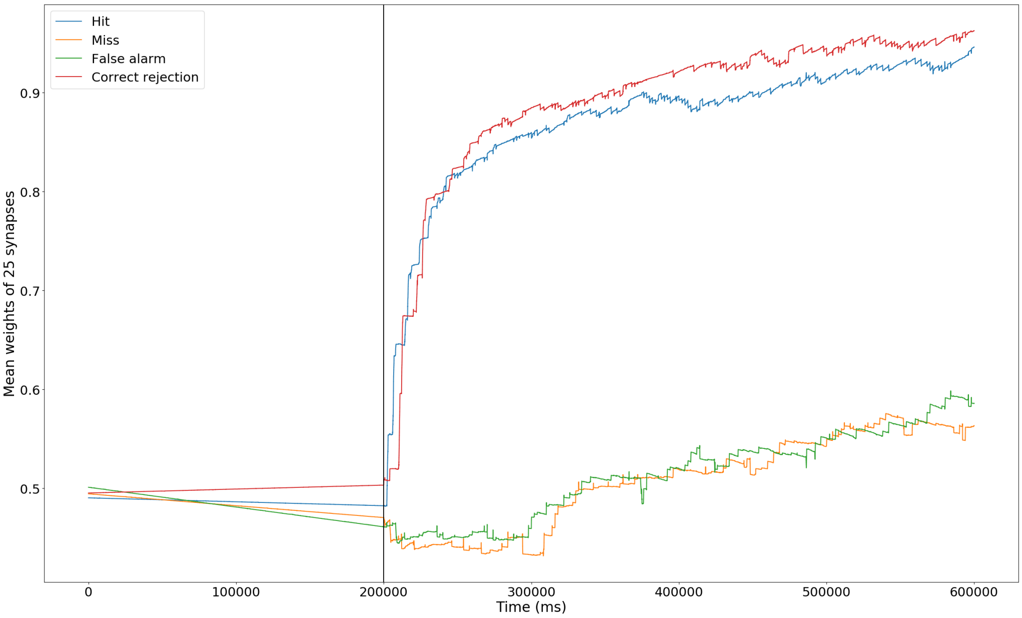


Figure 8: Evolution of synaptic weights connecting the middle neurons with the output neurons. The black line represent the moment of the transition from a naive state to a rewarded state. The mean of 25 weights was chosen to represent this synapses as the total number of synapses in the simulation was 600 and for computational performance reason not all of them could be monitored.

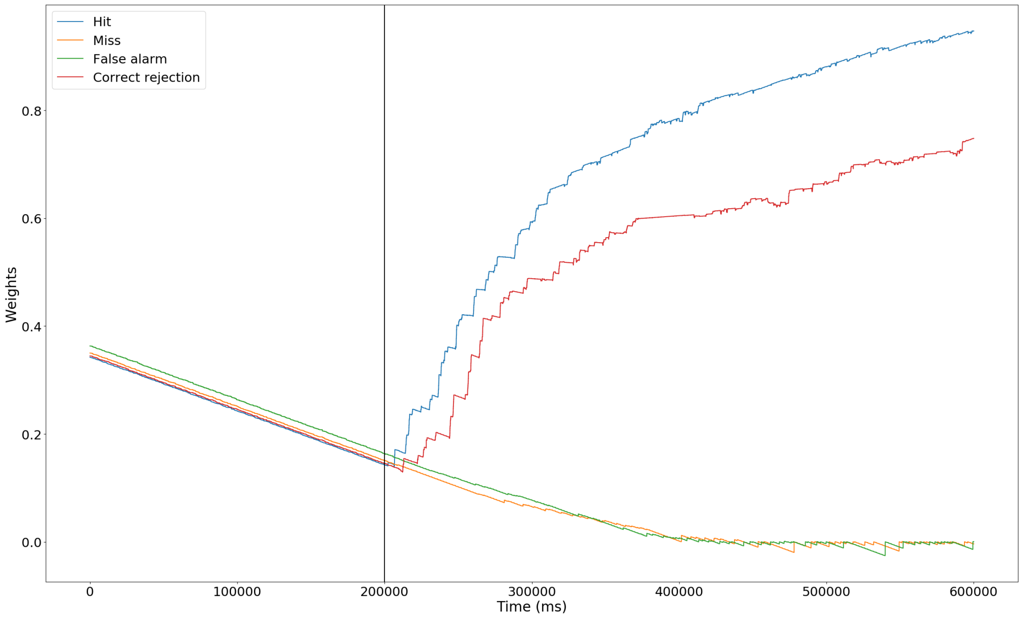


Figure 9: Evolution of synaptic mean weights connecting the four outcomes neurons with the critic neurons. The black line represent the moment of the transition from a naive state to a rewarded state.

Figure 10 shows that the firing rate of the critic is low when no reward is given. This is due to the absence of reward. When there is no reward, the reward neurons receive no input and thus they cannot excite the TD error neurons which in turn excite the critic and the only input the critic receives is the one coming from the four output association. This input however is not enough to make the critic fire significantly. When the reward signal is given the activity and firing of the critic gradually increases, as it gets more input and the synapses from middle to output are strengthened. At the same time TD error neuron activity rises as the reward signal activates an input current to the reward neurons which become active and thus excite the TD error neurons. However as mentioned before the weights of middle to output are low in the beginning of the simulation during the first times when reward is given and thus when the TD neurons are excited for the first time, the critic neurons have a low activity and firing rate. As a consequence the inhibition from the critic to TD error neurons is weak and the TD error activity is very high. At these times the learning is rate is high, because the high activity of the TD error that has been used to gate learning at the plastic synapses. While learning the activity of the critic gradually increases until it starts to correlate with the reward activity and the TD error activity goes down with time. After the model has learned the right association that is rewarded, a balance between inhibition and excitation of the TD error neurons is established, the critic fires synchronous to the reward neurons.

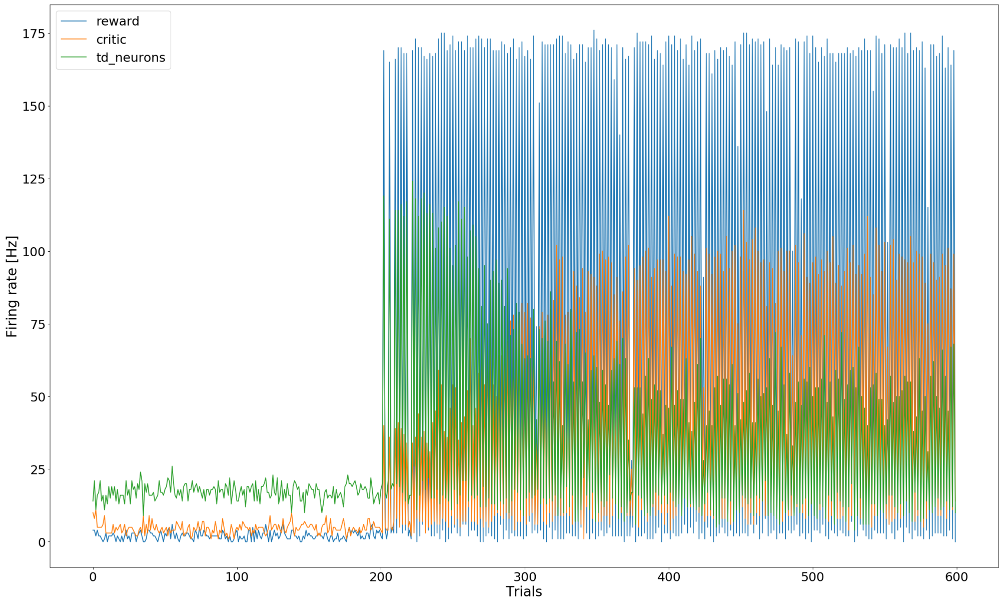


Figure 10: Evolution of the firing rate across different trials of the reward neurons, critic neurons and TD neurons.

Figure 11 shows the percentage of trials in which the different associations were chosen during the above showcased simulation. The mouse has chosen the correct associations during 40 percent of the trails and the non rewarded associations during around 20 percent of the trials.

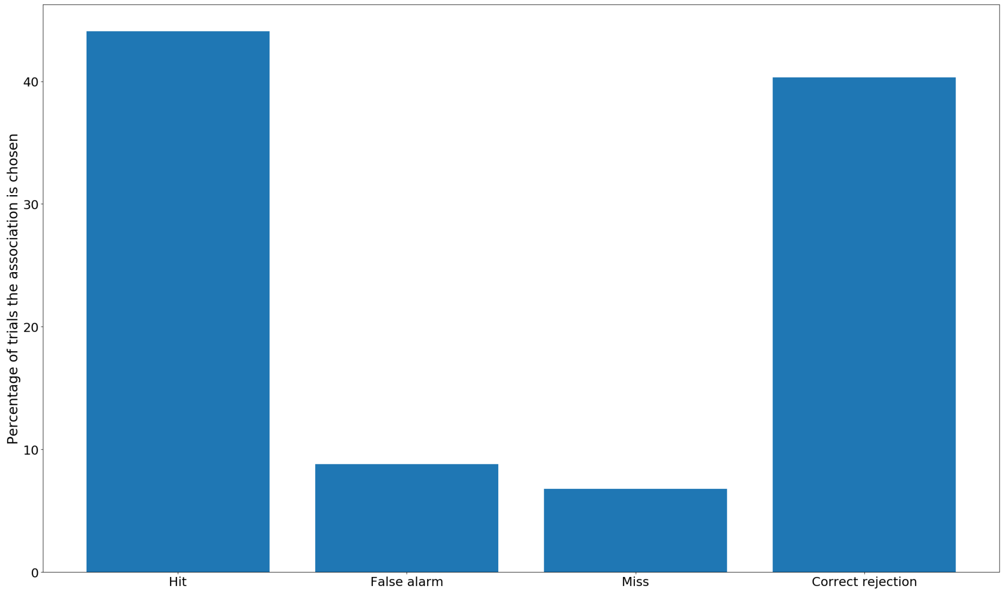


Figure 11: Bar chart showing the percentage of trials the four different behavioural outcomes have been chosen by the system.

3.3.2 Reversal

In some experiments the rewarded association is switched after some time. Results form these experiments show that mice learn faster the new rewarded association after the switching then the first rewarded association that has been learned. This switching has been incorporated in the model. Figure 12 shows the spiking pattern of a simulation where the first 200 seconds were unrewarded, followed by 200 seconds with reward and after the 400 trials the rewards are switched. This switch has as a consequence that the Hit becomes Miss and the Correct rejection a False alarm. Figure 12 and figure 13 show that after the first reward administration the systems learns quickly the right reward association and chooses often rewarded situation. The learning happens relatively quickly compared to the learning of the reversal. The synaptic weights after the reversal of the newly rewarded associations rise slower than the the prior rewarded association when it was learned. A reason for this is that as the previous rewarded association has high weights at the moment when the reward is switched and the system continues to choose those previously rewarded association. With time those weights go down as they are not rewarded any more and the newly association rises. Another reason is that the weights of the newly rewarded association start at a lower value then in the beginning of the simulation. As there is no negative learning incorporated in the system the old rewarded association are forgotten through extinction rather then actively. Experiments have shown that mice learn faster after the reversal of reward, which is partially due to the devaluation of the previously rewarded association. The absence of reward speeds up the forgetting, because it might emotionally impact the mouse. As the forgetting is faster the learning is also faster. Interestingly mice who are lesioned at the OFC show a similar behaviour as the one captured by the model. They fail to respond to devaluation and continue to learn at a similar rate or slower rate the new association compared to the previously learned association. The weights of the plastic synapses between the middle neurons and output neurons behave analogous to the figure 13 and thus are not shown here.

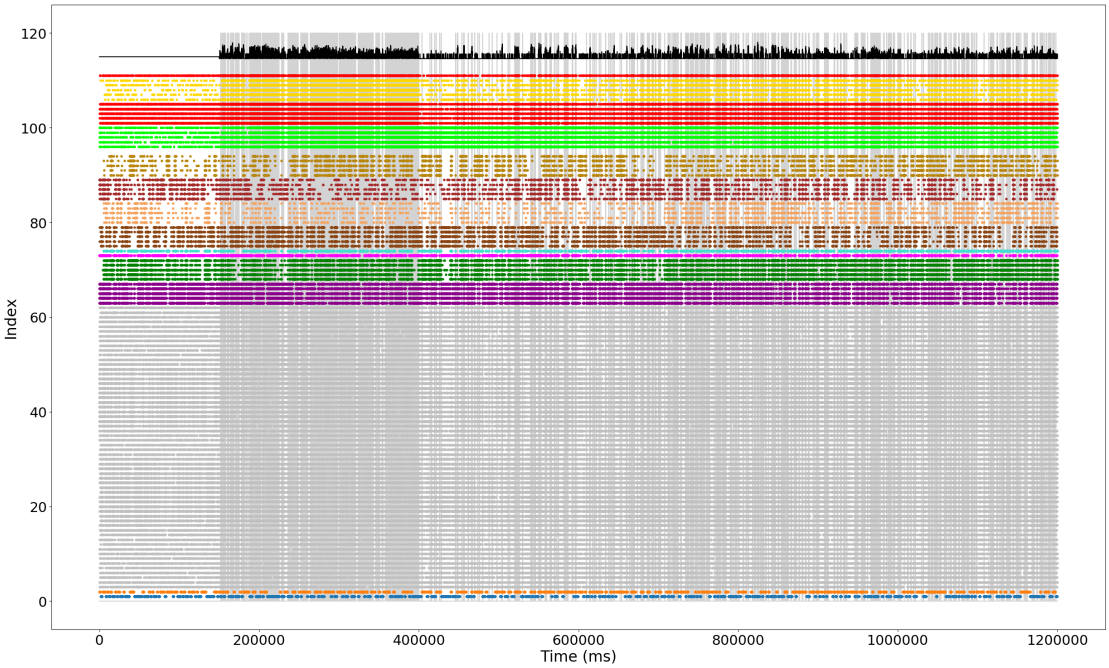


Figure 12: Spiking pattern of a simulation where the first 100 trials were unrewarded, followed by 100 rewarded trials and 400 trials in which the reward was switched. The light grey lines indicate when reward is given. Reward is given more frequently during the time when the first reward association is rewarded and less frequently after the switch.

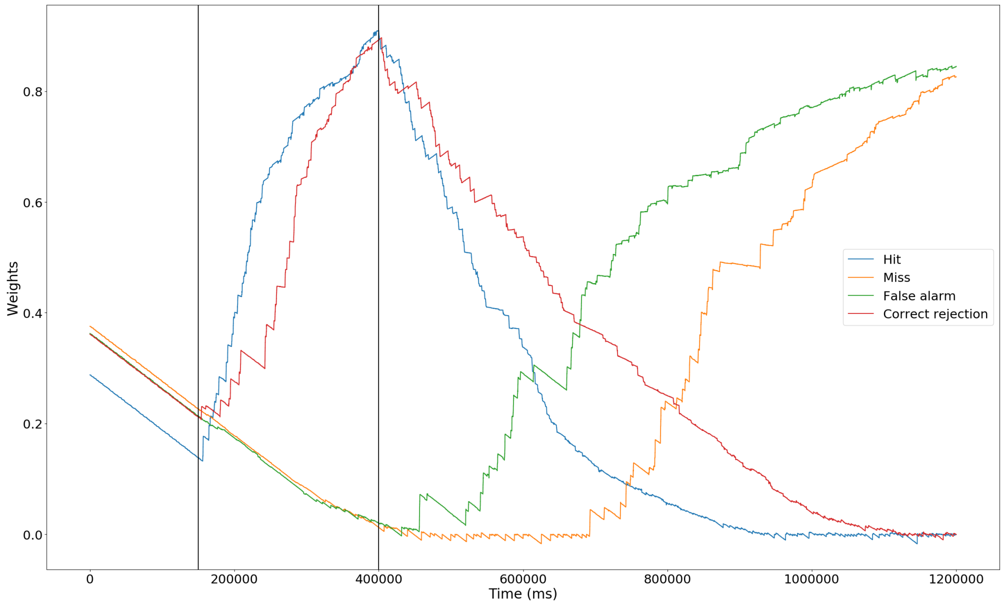


Figure 13: Evolution of synaptic mean weights from four outcomes neurons to the critic neurons. The black line represent the time when the system transitioned from a naive state to a rewarded state to a state that has reversed reward association.

4 Discussion

This section covers some debatable points of the model.

As mentioned in the architecture section the reward neurons receive input while the correct input and output association is active and in this time window learning is happening though TD error gating. In real life settings however reward is given after accomplishing a task and not while doing it and learning happens later. Therefore the reward should be given at the time after the trial has ended. This poses an implementation issue because the plastic learning rules follow the popular dictum ”Fire together, wire together”, thus the desired synaptic weights are only strengthened when neurons of the desired association fire. This cannot be assured if the reward is given after the end of the trial as the absence of input causes the system to return to baseline firing of neurons and noise could bias subsequently the association that is active at the delayed time and that association is likely to differ from the desired one.
A drawback of the model is the reversal learning, which has been quickly incorporated but not properly set up yet. In this simulation the association

of the input and output, which lead to reward was switched. After switching has happened the previous association needs to be unlearned while the new is learned. A certain rate of unlearning is necessary to induce the actor to do exploration. Experimental data of mouse brains also indicate that mice forget actively as they learn (Madroñal et al., 2016). This active forgetting is not captured by the model yet as their forgetting is induced by extinction. The synaptic weights go down because they synapses are not potentiating, while the other synapses of the newly rewarded association are. Furthermore learning the new association after the switch is currently slower, which seems to support experimental data of OFC lesioned mice but not wild type mice. Another possible drawback of the model is that all the trials have the same duration and in-between trials are pauses of equal duration to the trial. In the real life set-up however trial duration varies depending on the choice of the mouse as can be seen in figure 1. If the mouse licks for the non target texture it is punished by a time-out, which has as a consequence that the mouse needs to wait longer until it can claim the next reward. There is also no punishment for incorrect decisions in the computational model. Correct decisions are encouraged with reward but incorrect ones are forgotten by extinction because they don't yield reward and not because they are uncomfortable and need to be avoided. As a consequence after the switching of the associations the absence of reward does not lead to further devaluation of the previous rewarded association and the previous association is forgotten through extinction. Furthermore the emotional state of the mouse is not modelled. As mentioned above there is an absence of punishment and variable trial duration. Both of these might provoke an emotional response in the organism. A consequence of a variable trial duration is frustration because the mouse needs to wait longer for the reward and a consequence of punishment is aversion because of an uncomfortable feeling originating from it.

Furthermore, by examination of the architecture the reader might have noticed that there are neural populations that excite and inhibit other neurons at the same time and thus violate Dale's principle. Dale's principle in a nutshell consists of the idea that a neurons synaptic connections cannot be excitatory and inhibitory simultaneously (Osborne, 2013). This can be observed hower in this model, for instance in the output, binary, four outcomes and critic neurons (Figure 2). This choice has been made to save computation time and resources. The addition of further neuron groups performing inhibition would not have changed the overall functionality of the network. Additionally, Dale's principle has been critically reviewed in the past (Osborne, 1979; Sabelli et al., 1976). Experiments conducted in snails have indicated that serotonin neurons utilizes 5-Hydroxytryptamine (5-HT) and acetylcholine (ACh) as neurotransmitters (Kerkut, Sedden, & Walker, 1967; Emson & Fonnum, 1974; Cottrell, 1977). It is however debatable if this finding are evidence enough against Dale's law.

4.1 Possible localisation of model components in brain structures

In this section the different components of the model will be linked to what they could possibly represent in biology. This section is rather conceptual and should not be taken up as something absolute but rather ideas that are debatable.

The input neurons represent the sensory inputs. As the input neurons are generating input spikes to be propagated in the system, they could represent the nerve endings in the trigeminal ganglion that convert mechanical whisker energy into action potentials (AP).

The AP are then propagated to the middle group which could be seen as the somatosensory cortex. As they two stimuli which can be presented are not the same, they are thought to be processed by different neurons. This is modelled by the Gaussian connectivity of the synapses between input and middle group which connects differently the inputs to the neurons of the middle group depending on which input is active.

The output neurons could represent the motor area in the brain, which decides on the ouptut and the four outcomes neurons could be conceived as working memory, which holds the information about which input and output has been chosen until a reward is given and subsequently the TD error is computed. Therefore the four outcomes neurons might represent the OFC.

4.2 Future model improvement

The next step in this model is to modify the reversal learning to be more accurately describing the experimental data in wild type mice. In order to do so the synaptic weights of the previous learned association should be forgotten faster. An idea is to modify the plastic learning rule it in order to induce faster forgetting after switching rewards. Furthermore it would be interesting to look at existing learning and forgetting models (Jaber & Sikström, 2004) and to see if those could be incorporated into spiking neural networks and can be used to describe accurately experimentally observed reversal learning.

Another thing that would be interesting to implement in the model is punishment. Implementing this might help to solve the problem of reversal learning

because punishment could lead to devaluation.
Additionally the model needs to be fit to experimental conditions, by tuning the learning rate in order to learn at the same speed in the same time window mice learn.
Another possible modification of the model can be made to simplify the architecture. The four outcomes can be represented in the middle layer instead as an own neuron group. This can be achieved by making the middle group a 2D field that gets input from the inputs and has a recurrent connection from the output back into the field.

5 Conclusion

The model presented in this thesis provides a conceptual architecture with the attempt to capture the essence of learning in the mouse discrimination task by using ideas from reinforcement learning and spiking neural networks. In the beginning, that is in the absence of reward the system tries randomly association of input and outputs. As soon as reward is given the system learns the correct association. When this association is reversed the system keeps choosing the previously rewarded association. Thus this hinders exploration and slows down the learning of the new association. The learning is gated with a TD error signal that is generated by an interaction loop between TD error, reward and critic neurons and that is propagated to gate the Fusi learning rule at the plastic synapses. TD error neurons are excited when reward is given and excite the critic neurons, which in turn inhibit the TD neurons. Thus the TD neurons activity is determined by a balance from inhibition coming from critic neurons and excitation from TD error neurons. When this balance is established the learning rate is low but if there is an discrepancy in inhibition and excitation the learning is strong. This TD learning captures the learning well.The first time a system is presented with an reward association when this association between stimuli and outcome is switched the model fails to capture experimental observation. As the learned association keeps being selected and forgetting is slow. In experimental condition however the learning of an new association is faster. Interestingly this behaviour of the model, is similar to that seen in OFC lesioned animals.

6 Acknowledgement

Special thanks to Yulia Sandamirskaya for great support and for this interesting project and to Alpha Renner for help with the implementation. Further thanks to Héctor Vasquez Martinez for sharing his code and ideas, to the whole people of the Institute of Neuroinformatics for providing an inspirational working environment and finally big thanks to Fritjhof Helmchen and Daniel Kiper for supervision.

References

Arabzadeh, E., Panzeri, S., & Diamond, M. E. (2004). Whisker vibration information carried by rat barrel cortex neurons. *Journal of Neuroscience*, *24*(26), 6011–6020.

Arabzadeh, E., Panzeri, S., & Diamond, M. E. (2006). Deciphering the spike train of a sensory neuron: counts and temporal patterns in the rat whisker pathway. *Journal of Neuroscience*, *26*(36), 9216–9226.

Arabzadeh, E., Petersen, R. S., & Diamond, M. E. (2003). Encoding of whisker vibration by rat barrel cortex neurons: implications for texture discrimination. *Journal of Neuroscience*, *23*(27), 9146–9154.

Arabzadeh, E., Zorzin, E., & Diamond, M. E. (2005). Neuronal encoding of texture in the whisker sensory pathway. *PLoS biology*, *3*(1), e17.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4-5), 407–419.

Berg, R. W., & Kleinfeld, D. (2003). Rhythmic whisking by rat: retraction as well as protraction of the vibrissae is under active muscular control. *Journal of neurophysiology*, *89*(1), 104–117.

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain research reviews*, *28*(3), 309–369.

Boubenec, Y., Shulz, D. E., & Debrégeas, G. (2012). Whisker encoding of mechanical events during active tactile exploration. *Frontiers in behavioral neuroscience*, *6*, 74.

Boulougouris, V., Dalley, J. W., & Robbins, T. W. (2007). Effects of orbitofrontal, infralimbic and prelimbic cortical lesions on serial spatial reversal learning in the rat. *Behavioural brain research*, *179*(2), 219–228.

Brader, J. M., Senn, W., & Fusi, S. (2007). Learning real-world stimuli in a neural network with spike-driven synaptic dynamics. *Neural computation*, *19*(11), 2881–2912.

Brecht, M., Preilowski, B., & Merzenich, M. M. (1997). Functional architecture of the mystacial vibrissae. *Behavioural brain research*, *84*(1-2), 81–97.

Brecht, M., & Sakmann, B. (2002). Whisker maps of neuronal subclasses of the rat ventral posterior medial thalamus, identified by whole-cell voltage recording and morphological reconstruction. *The Journal of Physiology*, *538*(2), 495–515.

Brecht, M., Schneider, M., Sakmann, B., & Margrie, T. W. (2004). Whisker movements evoked by stimulation of single pyramidal cells in rat motor cortex. *Nature*, *427*(6976), 704.

Brown, V. J., & Bowman, E. M. (2002). Rodent models of prefrontal cortical function. *Trends in neurosciences*, *25*(7), 340–343.

Carvell, G. E., & Simons, D. J. (1990). Biometric analyses of vibrissal tactile discrimination in the rat. *Journal of Neuroscience*, *10*(8), 2638–2648.

Chakrabarti, S., Zhang, M., & Alloway, K. D. (2008). Mi neuronal responses to peripheral whisker stimulation: relationship to neuronal activity in si barrels and septa. *Journal of neurophysiology*, *100*(1), 50–63.

Chen, J. L., Carta, S., Soldado-Magraner, J., Schneider, B. L., & Helmchen, F. (2013). Behaviour-dependent recruitment of long-range projection neurons in somatosensory cortex. *Nature*, *499*(7458), 336.

Chen, J. L., Margolis, D. J., Stankov, A., Sumanovski, L. T., Schneider, B. L., & Helmchen, F. (2015). Pathway-specific reorganization of projection neurons in somatosensory cortex during learning. *Nature Neuroscience*, *18*(8), 1101.

Chen, J. L., Voigt, F. F., Javadzadeh, M., Krueppel, R., & Helmchen, F. (2016). Long-range population dynamics of anatomically defined neocortical networks. *Elife*, *5*.

Chudasama, Y., & Robbins, T. W. (2003). Dissociable contributions of the orbitofrontal and infralimbic cortex to pavlovian autoshaping and discrimination reversal learning: further evidence for the functional heterogeneity of the rodent frontal cortex. *Journal of Neuroscience*, *23*(25), 8771–8780.

Clarke, W., & Bowsher, D. (1962). Terminal distribution of primary afferent trigeminal fibers in the rat. *Experimental neurology*, *6*(5), 372–383.

Cottrell, G. (1977). Identified amine-containing neurones and their synaptic connexions. *Neuroscience*, *2*(1), 1–18.

Crochet, S., & Petersen, C. C. (2006). Correlating whisker behavior with membrane potential in barrel cortex of awake mice. *Nature neuroscience*, *9*(5), 608.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience* (Vol. 806). Cambridge, MA: MIT Press.

Deschênes, M., Timofeeva, E., Lavallée, P., & Dufresne, C. (2005). The vibrissal system as a model of thalamic operations. *Progress in brain research*, *149*, 31–40.

Diamond, M. E., Von Heimendahl, M., Knutsen, P. M., Kleinfeld, D., & Ahissar, E. (2008). 'where'and'what'in the whisker sensorimotor system. *Nature Reviews Neuroscience*, *9*(8), 601.

Dias, R., Robbins, T., & Roberts, A. (1996). Dissociation in prefrontal cortex of affective and attentional shifts. *Nature*, *380*(6569), 69.

Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, *22*(1), 1–18.

Dörfl, J. (1985). The innervation of the mystacial region of the white mouse: A topographical study. *Journal of anatomy*, *142*, 173.

Emson, P., & Fonnum, F. (1974). Choline acetyltransferase, acetylcholinesterase and aromatic l-amino acid decarboxylase in single identified nerve cell bodies from snail helix aspersa. *Journal of neurochemistry*, *22*(6), 1079–1088.

Fee, M. S., Mitra, P. P., & Kleinfeld, D. (1997). Central versus peripheral determinants of patterned spike activity in rat vibrissa cortex during whisking. *Journal of neurophysiology*, *78*(2), 1144–1149.

Feldmeyer, D., Brecht, M., Helmchen, F., Petersen, C. C., Poulet, J. F., Staiger, J. F., … Schwarz, C. (2013). Barrel cortex function. *Progress in neurobiology*, *103*, 3–27.

Fend, M., Yokoi, H., & Pfeifer, R. (2003). Optimal morphology of a biologically-inspired whisker array on an obstacle-avoiding robot. In *European conference on artificial life* (pp. 771–780).

Ferry, A. T., Lu, X.-C. M., & Price, J. L. (2000). Effects of excitotoxic lesions in the ventral striatopallidal–thalamocortical pathway on odor reversal learning: inability to extinguish an incorrect response. *Experimental Brain Research*, *131*(3), 320–335.

Frémaux, N., Sprekeler, H., & Gerstner, W. (2013). Reinforcement learning using a continuous time actor-critic framework with spiking neurons. *PLoS computational biology*, *9*(4), e1003024.

Friedberg, M. H., Lee, S. M., & Ebner, F. F. (1999). Modulation of receptive field properties of thalamic somatosensory neurons by the depth of anesthesia. *Journal of neurophysiology*, *81*(5), 2243–2252.

Gallagher, M., McMahan, R. W., & Schoenbaum, G. (1999). Orbitofrontal cortex and representation of incentive value in associative learning. *Journal of Neuroscience*, *19*(15), 6610–6614.

Gerdjikov, T. V., Bergner, C. G., & Schwarz, C. (2017). Global tactile coding in rat barrel cortex in the absence of local cues. *Cerebral Cortex*, 1–13.

Gerstner, W., Kempter, R., van Hemmen, J. L., & Wagner, H. (1996). A neuronal learning rule for sub-millisecond temporal coding. *Nature*, *383*(6595), 76.

Gerstner, W., Kistler, W. M., Naud, R., & Paninski, L. (2014). *Neuronal dynamics: From single neurons to networks and models of cognition.* Cambridge University Press.

Goldman-Rakic, P. S. (2011). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. *Comprehensive Physiology*, 373–417.

Goodman, D. F., & Brette, R. (2008). Brian: a simulator for spiking neural networks in python. *Frontiers in neuroinformatics*, *2*, 5.

Goodman, D. F., Stimberg, M., Yger, P., & Brette, R. (2014). Brian 2: neural simulations on a variety of computational hardware. *BMC neuroscience*, *15*(1), P199.

Gottfried, J. A., O'doherty, J., & Dolan, R. J. (2003). Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science*, *301*(5636), 1104–1107.

Guić-Robles, E., Valdivieso, C., & Guajardo, G. (1989). Rats can learn a roughness discrimination using only their vibrissal system. *Behavioural brain research*, *31*(3), 285–289.

Guo, Z. V., Hires, S. A., Li, N., O'Connor, D. H., Komiyama, T., Ophir, E., ... others (2014). Procedures for behavioral experiments in head-fixed mice. *PloS one*, *9*(2), e88678.

Helmchen, F., Gilad, A., & Chen, J. L. (2018). Neocortical dynamics during whisker-based sensory discrimination in head-restrained mice. *Neuroscience*, *368*, 57–69.

Hill, D. N., Bermejo, R., Zeigler, H. P., & Kleinfeld, D. (2008). Biomechanics of the vibrissa motor plant in rat: rhythmic whisking consists of triphasic neuromuscular activity. *Journal of Neuroscience*, *28*(13), 3438–3455.

Hodgkin, A. L., Huxley, A. F., & Katz, B. (1952). Measurement of current-voltage relations in the membrane of the giant axon of loligo. *The Journal of physiology*, *116*(4), 424–448.

Houk, J. C., Davis, J. L., & Beiser, D. G. (1995). *Models of information processing in the basal ganglia.* MIT press.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, *160*(1), 106–154.

Iversen, S. D., & Mishkin, M. (1970). Perseverative interference in monkeys following selective lesions of the inferior prefrontal convexity. *Experimental Brain Research*, *11*(4), 376–386.

Izquierdo, A., Suda, R. K., & Murray, E. A. (2004). Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. *Journal of Neuroscience*, *24*(34), 7540–7548.

Jaber, M. Y., & Sikström, S. (2004). A numerical comparison of three potential learning and forgetting models. *International Journal of Production Economics*, *92*(3), 281–294.

Jones, B., & Mishkin, M. (1972). Limbic lesions and the problem of stimulus—reinforcement associations. *Experimental neurology*, *36*(2), 362–

377.

Kempter, R., Gerstner, W., & Van Hemmen, J. L. (1999). Hebbian learning and spiking neurons. *Physical Review E*, *59*(4), 4498.

Kerkut, G., Sedden, C., & Walker, R. (1967). Uptake of dopa and 5-hydroxytryptophan by monoamine-forming neurones in the brain of helix aspersa. *Comparative biochemistry and physiology*, *23*(1), 159–162.

Kim, D., & Moeller, R. (2004). A biomimetic whisker for texture discrimination and distance estimation. *From animals to animats*, *8*, 140–149.

Kleinfeld, D., Sachdev, R. N., Merchant, L. M., Jarvis, M. R., & Ebner, F. F. (2002). Adaptive filtering of vibrissa input in motor cortex of rat. *Neuron*, *34*(6), 1021–1034.

Knutsen, P. M., Pietr, M., & Ahissar, E. (2006). Haptic object localization in the vibrissal system: behavior and performance. *Journal of Neuroscience*, *26*(33), 8451–8464.

Kolb, B., Nonneman, A. J., & Singh, R. (1974). Double dissociation of spatial impairments and perseveration following selective prefrontal lesions in rats. *Journal of comparative and physiological psychology*, *87*(4), 772.

Krupa, D. J., Matell, M. S., Brisben, A. J., Oliveira, L. M., & Nicolelis, M. A. (2001). Behavioral properties of the trigeminal somatosensory system in rats performing whisker-dependent tactile discriminations. *Journal of Neuroscience*, *21*(15), 5752–5763.

Madroñal, N., Delgado-García, J. M., Fernández-Guizán, A., Chatterjee, J., Köhn, M., Mattucci, C., ... others (2016). Rapid erasure of hippocampal memory following inhibition of dentate gyrus granule cells. *Nature communications*, *7*, 10923.

Mao, X. (2007). *Stochastic differential equations and applications*. Elsevier.

Marr, D. (1969). A theory of cerebellar cortex. *The Journal of physiology*, *202*(2), 437–470.

McAlonan, K., & Brown, V. J. (2003). Orbital prefrontal cortex mediates reversal learning and not attentional set shifting in the rat. *Behavioural brain research*, *146*(1-2), 97–103.

Mégevand, P., Troncoso, E., Quairiaux, C., Muller, D., Michel, C. M., & Kiss, J. Z. (2009). Long-term plasticity in mouse sensorimotor circuits after rhythmic whisker stimulation. *Journal of Neuroscience*, *29*(16), 5326–5335.

Mehta, S. B., & Kleinfeld, D. (2004). Frisking the whiskers: patterned sensory input in the rat vibrissa system. *Neuron*, *41*(2), 181–184.

Miyashita, E., Keller, A., & Asanuma, H. (1994). Input-output organization of the rat vibrissal motor cortex. *Experimental brain research*, *99*(2), 223–232.

Moore, C. I. (2004). Frequency-dependent processing in the vibrissa sensory system. *Journal of neurophysiology*, *91*(6), 2390–2399.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154.

O'Connor, D. H., Clack, N. G., Huber, D., Komiyama, T., Myers, E. W., & Svoboda, K. (2010). Vibrissa-based object localization in head-fixed mice. *Journal of Neuroscience*, *30*(5), 1947–1967.

O'Doherty, J. P., Deichmann, R., Critchley, H. D., & Dolan, R. J. (2002). Neural responses during anticipation of a primary taste reward. *Neuron*, *33*(5), 815–826.

Osborne, N. N. (1979). Is dale's principle valid? *Trends in Neurosciences*, *2*, 73–75.

Osborne, N. N. (2013). *Dale's principle and communication between neurones: Based on a colloquium of the neurochemical group of the biochemical society, held at oxford university, july 1982*. Elsevier.

Pammer, L., O'Connor, D. H., Hires, S. A., Clack, N. G., Huber, D., Myers, E. W., & Svoboda, K. (2013). The mechanical variables underlying object localization along the axis of the whisker. *Journal of Neuroscience*, *33*(16), 6726–6741.

Petersen, C. C. (2007). The functional organization of the barrel cortex. *Neuron*, *56*(2), 339–355.

Pickens, C. L., Saddoris, M. P., Setlow, B., Gallagher, M., Holland, P. C., & Schoenbaum, G. (2003). Different roles for orbitofrontal cortex and basolateral amygdala in a reinforcer devaluation task. *Journal of Neuroscience*, *23*(35), 11078–11084.

Polley, D. B., Rickert, J. L., & Frostig, R. D. (2005). Whisker-based discrimination of object orientation determined with a rapid training paradigm. *Neurobiology of learning and memory*, *83*(2), 134–142.

Prigg, T., Goldreich, D., Carvell, G. E., & Simons, D. J. (2002). Texture discrimination and unit recordings in the rat whisker/barrel system. *Physiology & behavior*, *77*(4-5), 671–675.

Rall, W. (1964). Theoretical significance of dendritic trees for neuronal input-output relations. *Neural theory and modeling*, 73–97.

Rolls, E. T., Hornak, J., Wade, D., & McGrath, J. (1994). Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. *Journal of Neurology, Neurosurgery & Psychiatry*, *57*(12), 1518–1524.

Sabelli, H., Mosnaim, A., Vazquez, A., Giardina, W., Borison, R., & Pedemonte, W. (1976). Biochemical plasticity of synaptic transmission: a critical review of dale's principle. *Biological psychiatry*, *11*(4), 481–524.

Schoenbaum, G., Chiba, A. A., & Gallagher, M. (1998). Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nature neuroscience*, *1*(2), 155.

Schoenbaum, G., & Roesch, M. (2005). Orbitofrontal cortex, associative learning, and expectancies. *Neuron*, *47*(5), 633–636.

Schoenbaum, G., Setlow, B., Saddoris, M. P., & Gallagher, M. (2003). Encoding predicted outcome and acquired value in orbitofrontal cortex during cue sampling depends upon input from basolateral amygdala. *Neuron*, *39*(5), 855–867.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of neurophysiology*, *80*(1), 1–27.

Sejnowski, T. J., Koch, C., & Churchland, P. S. (1988). Computational neuroscience. *Science*, *241*(4871), 1299–1306.

Seth, A. K., McKinstry, J. L., Edelman, G. M., & Krichmar, J. L. (2004). Spatiotemporal processing of whisker input supports texture discrimination by a brain-based device. In *Animals to animats 8: Proceedings of the eighth international conference on the simulation of adaptive behavior. meyer, ja (ed.) mit press, cambridge, ma* (pp. 130–139).

Shuler, M. G., Krupa, D. J., & Nicolelis, M. A. (2001). Bilateral integration of whisker information in the primary somatosensory cortex of rats. *Journal of Neuroscience*, *21*(14), 5251–5261.

Simons, D. J., & Carvell, G. E. (1989). Thalamocortical response transformation in the rat vibrissa/barrel system. *Journal of neurophysiology*, *61*(2), 311–330.

Skinner, B. F. (1935). Two types of conditioned reflex and a pseudo type. *The Journal of General Psychology*, *12*(1), 66–77.

Solomon, J. H., & Hartmann, M. J. (2006). Biomechanics: robotic whiskers used to sense features. *Nature*, *443*(7111), 525.

Stimberg, M., Goodman, D. F., Benichoux, V., & Brette, R. (2014). Equation-oriented specification of neural models for simulations. *Frontiers in Neuroinformatics*, *8*, 6.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, *3*(1), 9–44.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1) (No. 1). MIT press Cambridge.

Thorndike, E. L., et al. (1912). *Animal intelligence. experimental studies.* LWW.

Torvik, A. (1956). Afferent connections to the sensory trigeminal nuclei, the nucleus of the solitary tract and adjacent structures. an experimental study in the rat. *Journal of Comparative Neurology*, *106*(1), 51–141.

Trappenberg, T. (2009). *Fundamentals of computational neuroscience.* OUP Oxford.

Veinante, P., & Deschênes, M. (1999). Single-and multi-whisker channels in the ascending projections from the principal trigeminal nucleus in the rat. *Journal of Neuroscience*, *19*(12), 5085–5095.

Von Heimendahl, M., Itskov, P. M., Arabzadeh, E., & Diamond, M. E. (2007). Neuronal activity in rat barrel cortex underlying texture discrimination. *PLoS biology*, *5*(11), e305.

Wijaya, J. A., & Russell, R. A. (2002). Object exploration using whisker sensors. In *Australasian conf. on robotics and automation* (pp. 180–185).

Wolfe, J., Hill, D. N., Pahlavan, S., Drew, P. J., Kleinfeld, D., & Feldman, D. E. (2008). Texture coding in the rat whisker system: slip-stick versus differential resonance. *PLoS biology*, *6*(8), e215.

Woolsey, T. A., & der Loos Van, H. (1970). The structural organization of layer iv in the somatosensory region (si) of mouse cerebral cortex. the description of a cortical field composed of discrete cytoarchitectonic units. *Brain research*, *17*(2), 205–242.

Zuo, Y., Perkon, I., & Diamond, M. E. (2011). Whisking and whisker kinematics during a texture classification task. *Phil. Trans. R. Soc. B*, *366*(1581), 3058–3069.