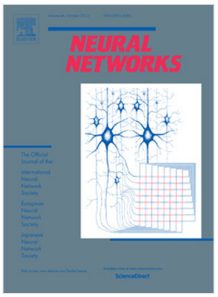
A model of operant learning based on chaotically varying synaptic strength

Accepted Manuscript



A model of operant learning based on chaotically varying synaptic strength

Tianqi Wei, Barbara Webb

PII: S0893-6080(18)30226-0
DOI: https://doi.org/10.1016/j.neunet.2018.08.006
Reference: NN 4010

To appear in: *Neural Networks*

Received date : 26 April 2018
Revised date : 12 July 2018
Accepted date : 2 August 2018

Please cite this article as: Wei, T., Webb, B., A model of operant learning based on chaotically varying synaptic strength. *Neural Networks* (2018), https://doi.org/10.1016/j.neunet.2018.08.006

A model of operant learning based on chaotically varying synaptic strength☆

Tianqi Wei[a,b,*], Barbara Webb[a]

[a]*School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom*
[b]*School of Engineering, University of Edinburgh, King's Buildings, Alexander Crum Brown Road, Edinburgh, EH9 3FF, United kingdom*

Abstract

Operant learning is learning based on reinforcement of behaviours. We propose a new hypothesis for operant learning at the single neuron level based on spontaneous fluctuations of synaptic strength caused by receptor dynamics. These fluctuations allow the neural system to explore a space of outputs. If the receptor dynamics are altered by a reinforcement signal the neural system settles to better states, i.e., to match the environmental dynamics that determine reward. Simulations show that this mechanism can support operant learning in a feed-forward neural circuit, a recurrent neural circuit, and a spiking neural circuit controlling an agent learning in a dynamic reward and punishment situation. We discuss how the new principle relates to existing learning rules and observed phenomena of short and long-term potentiation.

Keywords: Dynamic Synapse, Operant learning, Chaos, Receptor Trafficking

1. Introduction

Operant learning (also called operant conditioning or instrumental conditioning) is a type of learning in which a new behaviour is increased, or an existing behaviour is suppressed, by pairing it with reward or punishment. For example: (a) In a Skinner box, when a rat occasionally presses a lever, it

☆The work is funded by European Commission under FP7-ICT (Project ID: 618045).
*Corresponding author
Email address: chitianqilin@163.com (Tianqi Wei)

gets some food. After a while, it increases the rate of lever pressing (Jensen, 1963). (b) In a flight simulator, a fruit fly is heated when it generates yaw torque to one side and released from heat when it generates yaw torque to the other side. In minutes the fly learns to maintain its torque in the range that is without punishment (Wolf and Heisenberg, 1991) . (c) When an *Aplysia* produces a bite, the esophageal nerve can be stimulated in vivo to mimic the food signal. After training, it produces more bites than a yoked control that has received the same stimulation without the coupling to its own actions (Cash and Carew, 1989; Brembs, 2003).

Some of this research, e.g. in *Aplysia* (see review in Nargeot and Simmers (2011)), implies that mechanisms at the single neuron level can play important roles in operant learning. There are some existing single neuron or synapse models intended to account for operant learning. For example, the Hedonistic Synapse is a spike-based synapse model with stochastic synaptic transmissions, where the probability of transmitter release (the synaptic strength) is updated continuously according to the correlation between the transmitter fluctuation and a reward signal (Seung, 2003). Learning models based on modulated spike-timing-dependent plasticity (MSTDP) have also been applied to operant learning, using a reward signal to alter the weight of synapses that have been tagged by STDP as contributing to the output that produced the reward (for a review, see Frémaux et al. (2010)). These models only apply to spiking neural networks, and moreover, they have to introduce some arbitrary mechanism, such as a random number generator, to explore output space (i.e. generate different actions). Use of random number generators leads to the exploration of discrete output spaces with ever-present unpredictability.

An alternative option for generating exploration of the output space is chaos. Chaotic motion, which is a type of irregular motion that can exist in simple systems, has very complex, unpredictable and ergodic solutions (Tél et al., 2006; Eckmann and Ruelle, 1985). Chaos is widely found in biological systems (for a review, see Cavalieri and Koçak (1994)), including neurons and neural circuits. In a neuron, the dynamics of membrane potential and ion flows can be chaotic, as has been verified in several models, such as Nobukawa et al. (2014), Storace et al. (2008) and Canavier et al. (1990), and observed in the Nitella intermodal cell (Hayashi et al., 1983). Simulations of neural circuits also show chaos can exist at the circuit level, e.g. Sussillo (2014) and Angulo-Garcia and Torcini (2014). A chaotic system can be a source to generate unpredictable, continuous and ergodic actions for operant

learning or reinforcement learning. This idea has been applied to algorithms for robot learning, such as a Fish-Catching Robot that uses a chaotic generator for unpredictable motion planning to avoid fishes adapting to repetitive motions (Inukai et al., 2015) and a hexapod robot with a chaotic Central Pattern Generator (CPG) that produces chaotic signals for exploration of new motions to free its leg from a hole in the floor (Steingrube et al., 2011). The signals generated by a chaotic process are more continuous and more suitable for controlling a robots (or animals) interaction with the physical world than the signals generated by a random number generator, which are usually discrete white noise. Chaos in a physical system usually results in a more continuous and smooth variation of states than a random system. This property allows a transient delay of reward and modulator, which is common in learning in the real world. In principle, continuous and smooth trajectories can be obtained from a random number generator using interpolation, but, unlike chaos, the system will be predictable during the interpolation.

Although chaos is widely found in biological systems, the potential for chaos in synaptic dynamics and how this could support learning has not been previously considered. Here, we hypothesise that the following 'Dynamic Synapse' mechanism could underly operant learning (Fig 1). A neuron (Fig 1 (left)) has multiple input synapses, for which the synaptic strengths spontaneously fluctuate with uncorrelated phases (Fig 1 (right) green curve) around the centre of oscillation (Fig 1 (right) blue curve). We argue in more detail below that this could be caused by receptor trafficking. The neuron recieves inputs (e.g. from sensors or other neurons), and the inputs are multiplied by the synaptic strengths, summed up and passed through a non-linear function to determine the output. The output of the neuron causes some outcome (e.g. for an agent in an environment) which results in release of a neuromodulator according to a value function (Fig 1 (right) red curve). The modulator acts to bias the centre of the synaptic strength oscillation towards the instantaneous synaptic strength, and to decrease the amplitude of oscillation. Thus the synaptic strengths will converge to match the input-output properties of the neuron to the value function.

Is there a plausible biological mechanism that could produce the hypothesised synaptic strength fluctuation? The number of neurotransmitter receptors (from now on we will refer simply to receptors) embedded in the membrane of a post-synaptic dendritic spine is a key factor in synaptic strength (Sheng and Hoogenraad, 2007). Enlargement of a dendritic spine increases its capacity for anchoring structure, including scaffold proteins and cytoskele-

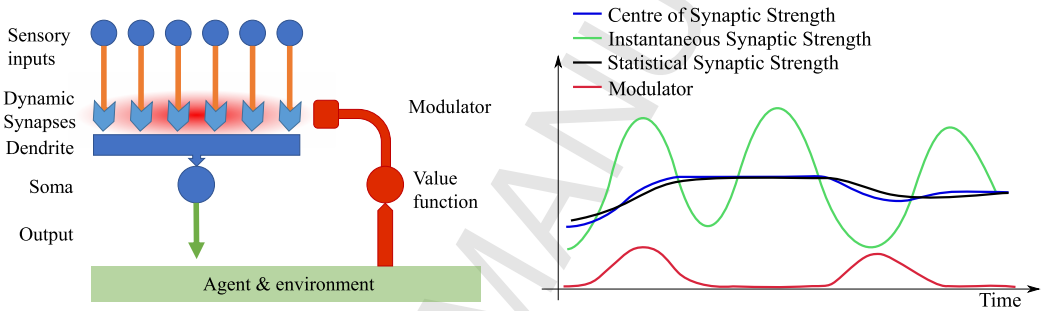


Figure 1: Basic concept of how operant learning works with a Dynamic Synapse. (Left): A neuron has multiple inputs, and its output is the sum of the inputs multiplied by the synaptic strengths, passed through a non-linear function. Because the synapses are dynamic, their values continuously change, and thus the output will explore a space of possible outputs. A value function on the output controls the release of a modulator which alters the synaptic strengths. (Right): Illustrating the dynamic synaptic strength of one synapse. During learning, the centre of synaptic strength oscillation is shifted towards the instantaneous synaptic strength that coincides with increased modulator, e.g., as illustrated, the modulator (red) is high when the instantaneous strength (green) is high, so the centre of synaptic strength is gradually increased (blue). The modulator also affects the damping of the oscillation, so the amplitude of oscillation decreases, and the learning can converge. An observer can infer the effective synaptic strength by low-pass filtering on the instantaneous synaptic strength (black) but note this is only an approximation of the actual centre of oscillation which cannot be directly observed.

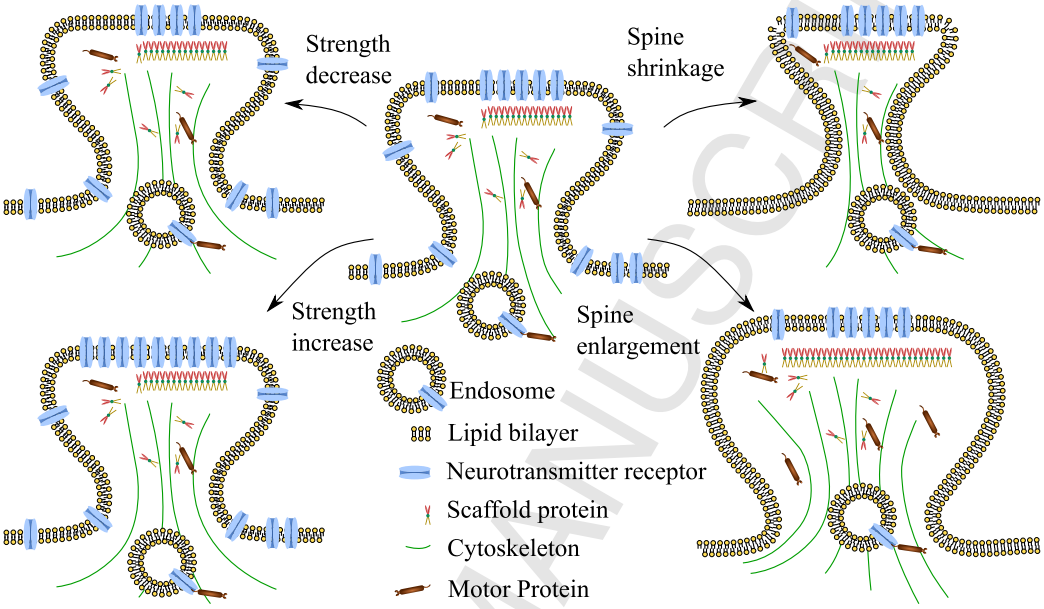


Figure 2: Decoupling between changes in spine size and synaptic strength under certain conditions. The membrane is formed mainly by the lipid bilayer and proteins. Cytoskeleton supports the shape of the dendrite spine. There are two forms of receptor trafficking. Lateral movement of receptors is observed as Brownian motion on the membrane. Endosomal trafficking carries receptors driven by motor protein along the cytoskeleton. Scaffold proteins can help receptors to anchor, increasing the capacity of the dendrite spine to hold the receptors. On the left, the size of neural spine stays the same, but the synaptic strength (number of receptors) varies. On the right, the size of dendrite spine varies, but the synaptic strength stays the same. Modified from Cingolani and Goda (2008)

ton, and thus the number of neurotransmitter receptors it can accommodate (Allison et al., 1998). However, the size and the capacity are not closely coupled (Cingolani and Goda, 2008). As shown in Fig 2, under certain conditions, synaptic strength can change without changes in spine size, and spine size can change without changes in synaptic strength.

The number of receptors in the membrane of a spine is also affected by two broad types of movement between synaptic and non-synaptic pools: lateral movement, which is mainly passive diffusion on the cell membrane; and endosomal trafficking, which is active transportation (Lau and Zukin, 2007). The lateral movement is affected by the cytoskeleton, which restricts or guides the diffusion (Jaqaman et al., 2011). In particular, the actin cytoskeleton has an active contribution to the regulation of postsynaptic receptor mobility both

in and out of synapses (Cingolani and Goda, 2008). The endosomal trafficking includes endocytosis of receptors from cell membrane to endosome, intracellular transportation of endosome, and exocytosis of receptors from endosome to the cell membrane (Roth et al., 2017). Endosomal trafficking can recycle receptors, transporting them between different regions (Petrini et al., 2009). There are also ongoing processes of receptor synthesis and degradation (Triller and Choquet, 2005).

The timescale of these receptor dynamics can be relatively fast. Receptors move from synaptic to extrasynaptic regions and vice versa usually with periods of up to a few minutes (Triller and Choquet, 2005). The size of a post-synaptic dendrite spine and the amount of actins in it oscillate in a time scale from tens of seconds (in immature dendrite spine) to a half hour (in a mature synapse)(Koskinen and Hotulainen, 2014; Honkura et al., 2008). Receptors anchored to the actin cytoskeleton (Hausrat et al., 2015) can move with the actin flow (Sergé et al., 2003). Post-synaptic receptor dynamics have been modelled at a mesoscopic level treating the regulation of numbers of the receptors and scaffold proteins as quasi-equilibrium based on thermodynamic theory (Sekimoto and Triller, 2009). The model proposed in Haselwandter et al. (2011) describes formation and stability of synaptic receptor domains as a reaction-diffusion system. We note these models are dynamic, but not chaotic. We propose i) that the complexity of post-synaptic dynamics (Choquet and Triller, 2013), especially receptor trafficking (Triller and Choquet, 2005) can support chaos and ii) that this can provide a mechanism for operant learning as described in Fig 1.

It is notable that dopamine has been shown to affect the same receptor trafficking dynamics (Sun et al., 2008). This supports the possibility that, in an operant learning paradigm, the relationship between the current synaptic strength (changing chaotically due to receptor trafficking) and a reward (signalled by neurotransmitter release) is a basis for learning.The possible role of alteration in postsynaptic receptor distribution and size of dendritic spines in learning (particularly in short-term and long-term potentiation (STP & LTP) protocols) is well established (Isaac et al., 1995; Kauer et al., 1988; Shepherd and Huganir, 2007). In Shouval et al. (2002), Shouval et al. proposed a thermodynamic model of AMPA receptor endosomal trafficking to explain bi-directional synaptic strength variation during LTP and long-term depression (LTD). Xie et al. (1997) proposed a synapse level model in which AMPA receptors are attracted toward NMDA receptors during STP, and some of the AMPA receptors become anchored near the NMDA receptors while others

diffuse again during LTP. The plausibility that such changes in receptor distribution could alter synaptic efficiency has also been demonstrated (Allam et al., 2015).

In the learning model presented here, we do not include any Hebbian process (see discussion). Instead, we allow chaotic synapses in a neuron to explore possible synaptic strengths; the neuron thus becomes a function on its inputs with chaotic coefficients, generating unpredictable output signals to explore action spaces. If the consequences of the action are reflected in a reinforcement signal delivered to the synapses, the parameters of the chaos can be altered to centre around synaptic strengths that optimise the output. We show through simulation the learning functionality of such a system in several different scenarios.

2. Result

Our model simplifies the structure of a neuron to consist of multiple input synapses and a dendrite, which together comprise the dendritic tree (Fig 3). We do not model the soma and axon of the neuron but simply calculate the somas input as the sum (across the dendritic tree) of the synaptic inputs multiplied by their respective synaptic strengths, then calculate the somas output by passing the input through a non-linear function. The number of receptors in a synapse represents the synaptic strength of the synapse. Receptors in the dendrite do not contribute any synaptic strength. Because of the receptor trafficking dynamics, the synaptic strength fluctuates spontaneously. In the methods we provide an abstracted mathematical model for receptor trafficking, but summarise here the key properties needed to support learning:

1. Spontaneously and smoothly varying synaptic strength w_i around an oscillation centre w_{ci};
2. The phases of the oscillations are not locked
3. The oscillation centre w_{ci} and amplitude depend on properties of the dendrite tree that can be altered by a learning signal.

When a neuron or network of neurons with such synapses produces output in a way that meets a specific requirement (given by a value function), modulator representing reward is released. The modulator affects the centre of synaptic strength oscillation, which shifts towards the instantaneous synaptic strength at the time of the modulator release. The simplest way

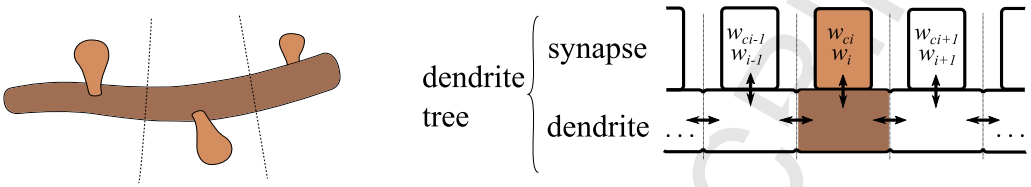


Figure 3: (Left) A dendrite tree consists of a dendrite (in dark brown) and multiple synapses (in light brown). (Right) A schematic diagram of the dendrite tree. Receptors can move between dendrite and synapse to dynamically modify the synapse strength w_i around some centre W_{ci}.

to implement this is as a learning rule depends only on the current centre of synaptic strength oscillation, the instantaneous synaptic strength and amount of the modulator:

$$\dot{w}_{ci} = \mathrm{k}_w(w_i - w_{ci})n_{\mathrm{M}} \tag{1}$$

where n_M is amount of the modulator, and k_w is a coefficient controlling the learning rate. By this learning rule, a circuit with dynamic synapses can conduct operant learning, as the instantaneous synaptic strength is near or in the range that satisfy a criterion when modulator is released (note in the experiments that follow we use a slightly altered rule (equation 23 in Methods) to compensate for a biased drift in synaptic strength). To allow learning to converge, the learning rule should also reduce the oscillation amplitude (equation 24). Conceptually, we relate the centre of oscillation to the capacity of a dendritic spine to hold receptors (Fig 2; and the amplitude of oscillation to the damping of the receptor movement dynamics. We assume these can result from changes in spine size or to the scaffold cyto-skeleton complex, but do not model these explicitly.

2.1. Simulation of a dendrite tree

In Fig 4, we show in simulation that our receptor trafficking model produces apparently chaotic and unpredictable oscillation of the synaptic weights. The simulated dynamic synapse system has six synapses, and the trajectory of the first three is plotted: it can be seen that it samples relatively evenly in the space of synaptic weight values. Fig 4 (right) shows how the range of exploration can be controlled. If the damping factor of a synapse increases, oscillation in the corresponding dimension of the plot will be narrower. If the capacity of a synapse changes, the centre of oscillation of the corresponding

dimension in the plot will translate. These properties are the basis of the principle by which the system can learn and converge. In this example, the periods of the oscillations are from 10 s to 20 s. With different parameters, the periods can be in a different range, such as in tens of minutes or hours, and the oscillations still appear chaotic after the equivalent of several days of simulated time. It is important for learning in our model that the synaptic dynamic timescale matches the causal dynamics of the learning situation. That is, when the reward is delivered, the state of the synapse should still be near the state that caused the action that resulted in reward. However, the timescale cannot be too long or else the generation of new actions will be limited, and the learning might converge to a local minimum. We note there may be other factors that produce unpredictable synaptic strengths, such as Brownian movement of receptors due to thermal noise, but suggest that these may be subsumed within the higher level dynamics described above, and it is not necessary to include them as a source of noise to support learning.

2.2. Applying learning in a simple linear example

In this experiment we test learning in a single neuron with reward provided when the output is higher than a threshold and increasing. The neuron is a linear neuron, i.e. its output is the sum of the product of input values and their synaptic strengths. During the simulation, the input values of the neuron are constants ranging from 0-5 as shown in Fig 5. The reward function is:

$$n_{\mathrm{m}} = \begin{cases} \mathrm{k_{m_1}} \dot{y}(y - y_0) & \text{if } \dot{y} > 0 \ \wedge \ y - y_0 > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where n_m is the amount of modulator, k_{m_1} a coefficient, y the output of the neuron, and y_0 a threshold of y to trigger the release of modulator.

Fig 6 (a) shows the instantaneous synaptic strengths, and the labels of lines show the constant input value of corresponding synapses. The equilibrium synaptic strengths, which are also average synaptic strengths, are shown in Fig 6 (b). Note that the later equilibrium synaptic strengths have the same ordering from highest to lowest as input strengths. The neuron has a fixed total of receptors, for which it finds an efficient distribution across the synapses to maximise. Fig 6 (c) shows the output of the neuron. In the first half of the learning process, the output decreased a little because the initial value is high but not stable. In the second half, the output gradually

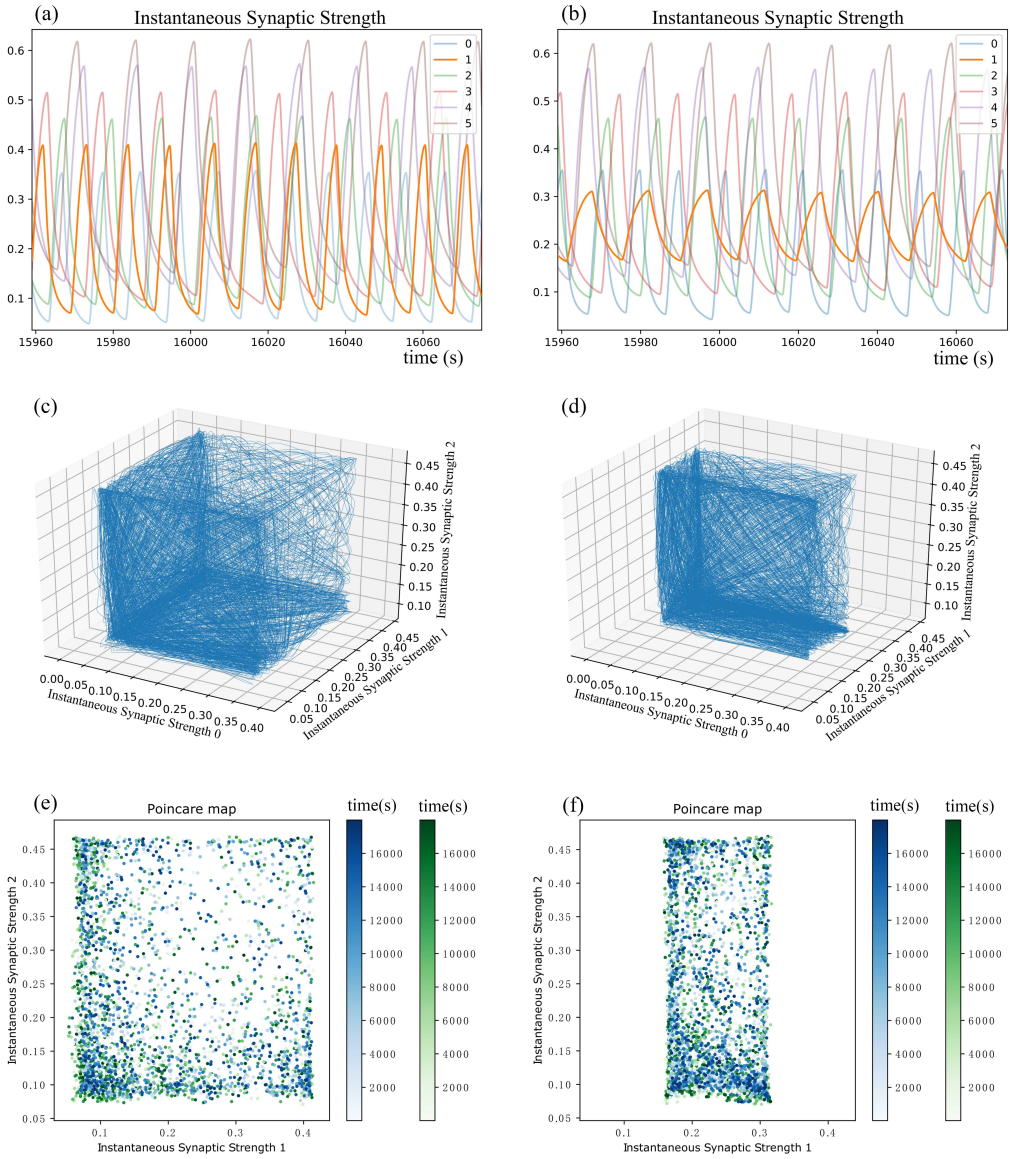


Figure 4: Trajectories of synaptic strengths. (Left): all synapses have the same damping factors. (Right): synapse one has a higher damping factor than others. (a) & (b) show the change over time of the synaptic strengths (the proportional number of receptors in each synapse); (c) & (d) plot the trajectory formed by the first three synapses (for (d) the synapse on the X-axis has higher damping); (e) & (f) are Poincaré maps, i.e., sections of (c) and (d) when the instantaneous synaptic strength passes the plane defined by the centre of oscillation for one synapse (blue and green are for two different directions, and time of intersection is indicated by the intensity). It can be seen that synaptic strength oscillates chaotically and unpredictably, tracing out a search space. With higher damping factors, the amplitude of the oscillation for that synapse is decreased, reducing the search space. The periods of the oscillations can be different with different parameters.

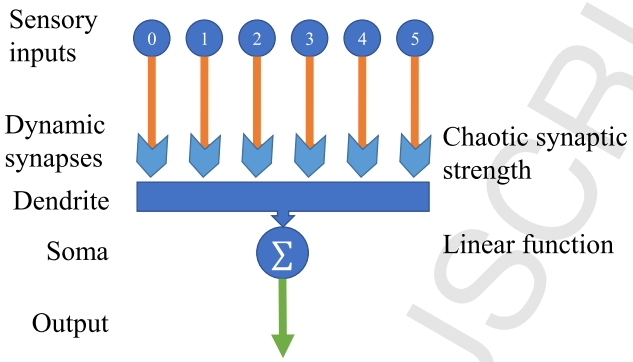


Figure 5: A linear neuron with dynamic synapses and several constant inputs. Its output is the sum of the inputs, each weighted by the respective synaptic strength.

increased. Fig 6 (d) shows the trajectory of first three synaptic strengths. The trajectory starts by exploring a large volume then gradually converges.

2.3. Tuning the period of a central pattern generator

A Central Pattern Generator (CPG) is a type of Recurrent Neural Network (RNN) which exists in many animals to control rhythmic motions, such as walking and heartbeat. It is also applied in legged robot control as an alternative to explicit motion planning (Ijspeert, 2008; Xia et al., 2017). However, online training of a CPG is difficult. People often have to tune it by hand or by offline parameter optimisation, such as brute force search or Genetic Algorithms. Our approach has a potential advantage in tuning or training a CPG because it can train a CPG online. This experiment shows an example of tuning a CPG to change its period. The CPG model is modified from the model described in Mori et al. (2004). The CPG is symmetric, and the synapses are replaced by Dynamic Synapses (as shown in Fig 7). The initial values of dynamic synaptic strengths were set to be the original synaptic strengths, and the initial amplitude of oscillation of synaptic strengths are scaled by an exponential function to be in the nearby order of magnitude of the original synaptic strengths.

$$w_{i_{cpg}} = w_{i_0}\beta^{w_i - 0.5} \tag{3}$$

where $w_{i_{CPG}}$ is CPG synapses weights, w_{i_0} the ith initial synaptic weight of the CPG, β is a base of exponentiation that scales the weights. As the CPG is symmetric, in the model, the state of dynamic synapses of one neuron is

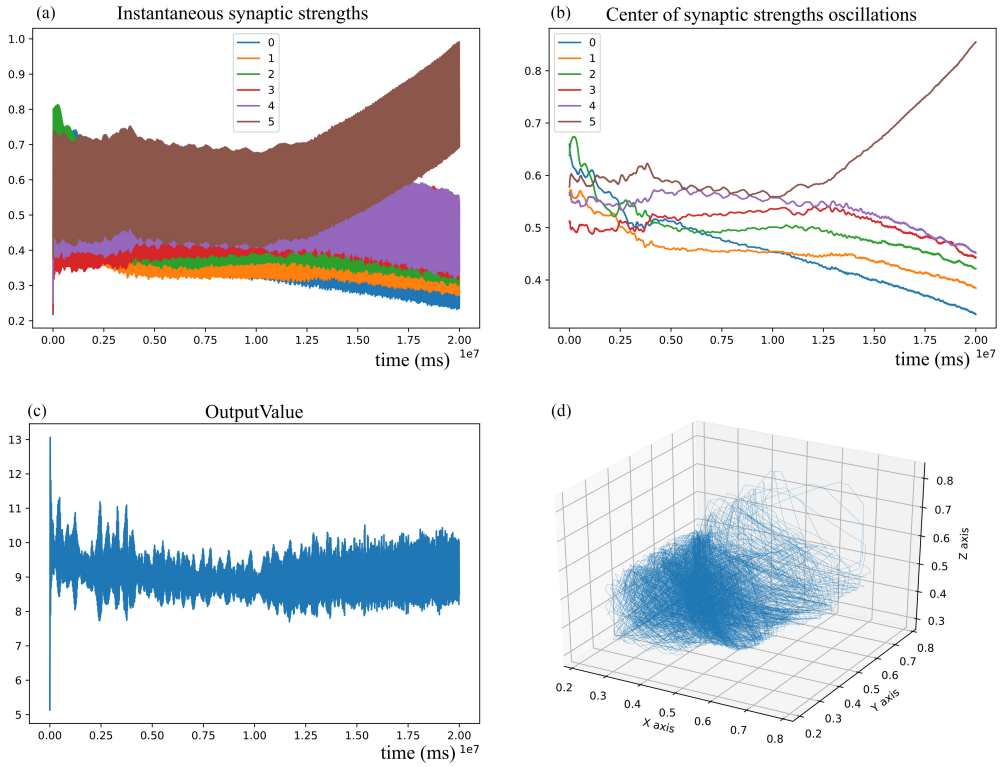


Figure 6: Simulation results of the simple linear example. The value function determining modulator release is that the output is higher than a threshold and increasing. (a) The instantaneous synaptic strengths, the labels of lines show the input value of corresponding synapses (b) the central synaptic strengths (c) the output value of the neuron (d) trajectory of the first three synaptic strengths. Note that the statistical output value starts to increase after unstable initial fluctuation. At the end of the learning, the centre of the oscillation of the synaptic strength shifts so that the order of strengths is the same as the order of the input values, and the synaptic strength of the synapse with highest input value increased while the others declined, which is the most efficient way to get higher output with conservation of the total number of receptors.

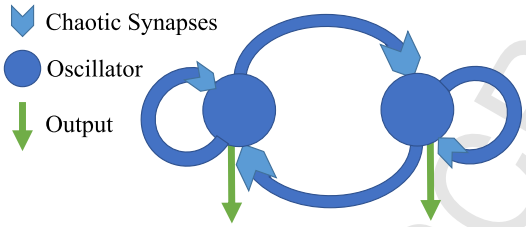


Figure 7: A CPG with the learning rule. Two neurons with spontaneous firing inhibit each other's firing alternately. The simulation aims to tune the period of oscillation, using the same operant learning rule to alter the synaptic strengths.

a mirror of the other one. When the output of the CPG crosses zero, the error between the target period and the actual period is calculated, and the modulator is released at a speed that is proportional to the decline of the error compared with the previous error. If the error increased, no modulator is released:

$$\epsilon_i = \omega_i - \omega_{obj} \tag{4}$$

$$n_{\mathrm{m_i}} = \begin{cases} \mathrm{k_{m_2}}(|\epsilon_{i-1}| - |\epsilon|) & \text{if} \quad |\epsilon_{i-1}| - |\epsilon| > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where ω_i is the period of the CPG from ith to $i+1$ th zero crossing, ω_{obj} the target period, ϵ_i is the error between them, n_{m_I} the amount of modulator released.

The CPG originally had a period of about 0.5 seconds. The target of training is to alter the period to be 2 seconds by tuning the synaptic strengths. The results are shown in Fig 8. Using the same operant learning rule as before, the period of the CPG converges to the target period. The period of the output of CPG and the synaptic strength is nonlinear and dynamic synapses have no prior knowledge of the CPG, but the simple neural circuit still finds and learns the parameters of the target effectively. The experiment shows that the Dynamic Synapse can be applied to an RNN without requiring any specific analysis of the properties of the network.

2.4. Reinforcement learning in Puckworld

The Dynamic Synapse model was tested in a game named PuckWorld, available as part of the Python Learning Environment. The game has a planar environment with three agents (Fig 9): a player that is controlled by a reinforcement learning algorithm, a reward source that changes its location

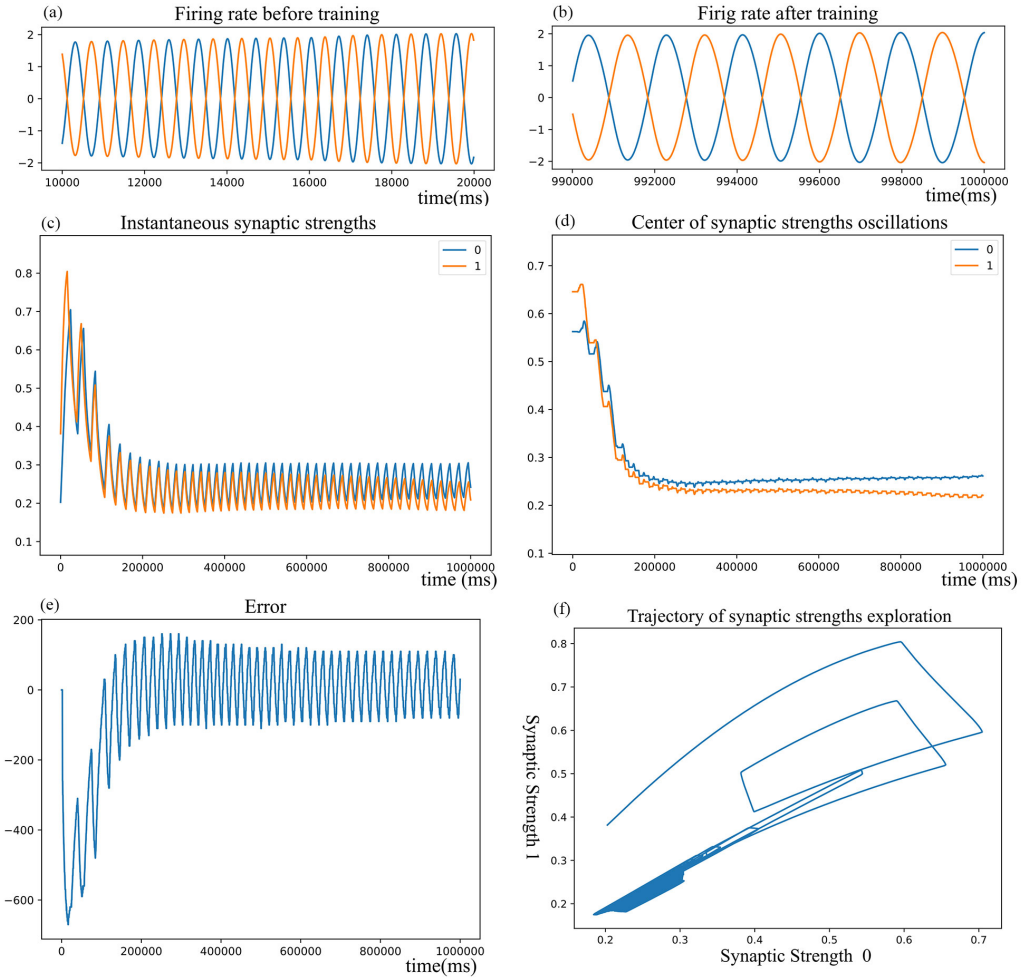


Figure 8: Results of tuning CPG with Dynamic Synapse. (a) Before learning the period of oscillation is about 500ms. (b) After learning the period of oscillation is about 2000ms. (c) The instantaneous synaptic strengths before scaling by the exponential function. As the model is symmetric, the two neurons share same states of dynamic synapses. Hence, only two synapses are plotted. Same in (d) and (e). (d) The centre of synaptic strength oscillation before scaling by the exponential function. (e) The error between the period of the output of the CPG and the target period during simulation. (f) the trajectory of chaotic exploration of the synaptic strength, which converged on the bottom left.

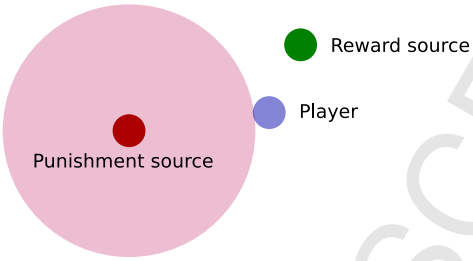


Figure 9: The environment of PuckWorld. The green point is the reward source, the blue point is the player, the red point is the punishment source, and the dark magenta circle is the range the punishment source effects.

after a specific period, and a punishment source that chases the player and decreases the reward if the player is within a specific range of the punishment source.

In the game, the player can move in 4 directions: left, right, down and up. The states of the player and the environment can be observed (Fig 10). The states are the velocity of the player, the locations of the player, the position of the reward source and the position of the punishment source. The states are pre-processed then used as sensor input. In this instance, the sensory inputs are the velocity of the player, the distance to the reward source, and the shortest distance the player is from the edge of the range of the punishment source (the distance to escape). As the game codes the states using an absolute coordinate system, the player does not have orientation. To transform the potentially negative values and direction of distance information in absolute coordinates into positive sensor values, the player is assumed to have sensors in 4 directions that correspond to the positive and negative directions of the x- and y-axis of the coordinate system, and the sensor on the side of the agent information coming from is positive, while the other side is zero (Fig. 10). As the player has a symmetric structure, the neural circuits are designed in a symmetric structure: four integrate-and-fire motor neurons control the motion in the four directions, respectively. Each neuron gets three types of sensory inputs (as outlined above) in the four directions. Each sensory input feeds into the neuron through a dynamic synapse. Also because of the symmetry of the structures and motions, to simplify and accelerate the training, the dynamic synapses of each motor neuron from sensors in the same direction relative to that motor neuron are treated as the same (have the same dynamics and parameters during the learning).

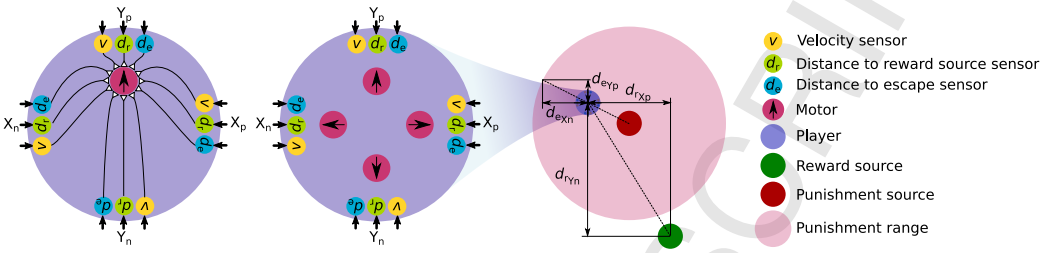


Figure 10: Sensors and neural circuits model for PuckWorld. (a) Velocity (v) sensors, distance to reward source (d_r) sensors and distance to escape (d_e) sensors get input from four directions; a motor neuron gets all of the sensory inputs by Dynamic Synapses. (b) There are four sets of neural circuits in the player; because the neural circuits, agents and the environment are symmetric, all homologous synapses are assumed to share the same dynamics and synaptic strengths to accelerate the learning. (c) The sensors indicate distances by orthogonal decomposition; when a measured object is in the direction that can be projected to the positive direction of a sensor, the sensory value is positive, otherwise 0.

The function of the motor neurons is:

$$\dot{\mathrm{v}} = \sum_{i=1}^{n} w_i s_i \tag{6}$$

$$\text{if } \mathrm{v} > \mathrm{v}_{\text{threshold}} \quad \mathrm{v} = \mathrm{v}_{\text{rest}} \tag{7}$$

where v is membrane potential, s_i the ith sensory input, v_{rest} the rest membrane potential and $v_{threshold}$ the threshold of firing.

The reward of the game is the weighted sum of the normalised distance to the reward source and the normalised distance into the range of the punishment source:

$$R = \begin{cases} -(d_r + 2d_e) & \text{if player is in punishment range} \\ -d_r & \text{otherwise} \end{cases} \tag{8}$$

where R is reward, d_r the distance between player and reward source, d_e the distance between player and the edge of punishment range.

The reward is fed into a firing rate neuron with an adaptive current, which releases the modulator. With the adaptive current, the neuron is sensitive to the change of the reward but insensitive to the value of the reward. The adaptation speed factor from low to high is higher than the

adaption speed factor from high to low, thus the neuron has a trend to increase the expectation of the reward:

$$\dot{I}_{adapt} = \begin{cases} (k_r R + I_{adapt})\, k_{adapt_1} & \text{if } R > I_{adapt} \\ (k_r R + I_{adapt})\, k_{adapt_2} & \text{if } R < I_{adapt} \end{cases} \tag{9}$$

where I_{adapt} is the current intensity, k_R a factor from reward to current intensity, k_{adapt_1} and k_{adapt_2} are factors of adaption speed. Thus modulator amount n_m is given by:

$$n_m = 2/(1 + e{-}k_{mI}(k_R R - I_{adapt}) - 1 \tag{10}$$

where k_{mI} is a factor to map the current after adaption to an appropriate range.

As this is a single layer circuit, the ability of a player controlled by the circuit is simple and limited. Hence, we can analyse the possible best solution of the synaptic strengths and compare it with the solution obtained by operant training with dynamic synapses. Treating the single layer circuit as a linear function, the whole system can be interpreted as a second-order system. For an appropriate solution, the interactions of the elements in the system should work as though (1) there is an extension spring connecting the player and reward source; (2) the punishment range is an elastic ball that pushes the player away; and (3) the elastic coefficient of the elastic ball is higher than the elastic coefficient of the spring so the player will avoid punishment even when the reward is inside the punishment range. Because of (1), the synaptic strengths of positive y distance to reward input should be higher than the synaptic strengths of negative y distance to reward input; because of (2), the synaptic strengths of positive y distance to escape input should be higher than the synaptic strengths of negative y distance to escape input; and because of (3) the synaptic strengths of positive escape input should be higher than the synaptic strengths of positive reward input.

The simulation results are shown in Fig 11. The simulation result was largely consistent with the analysis above, as shown in Fig 11 (a) and (c). However, surprisingly the highest synaptic strength is for negative x distance to reward input (line 4 in Fig 11 (a)) are higher than other lines, which means the agent would go forward when the reward source is on its left side. The positive y velocity (line 3) is also higher than negative y velocity (line 2), which means the agent tends to accelerate. These appear to be two strategies to avoid chasing by the punishment source.

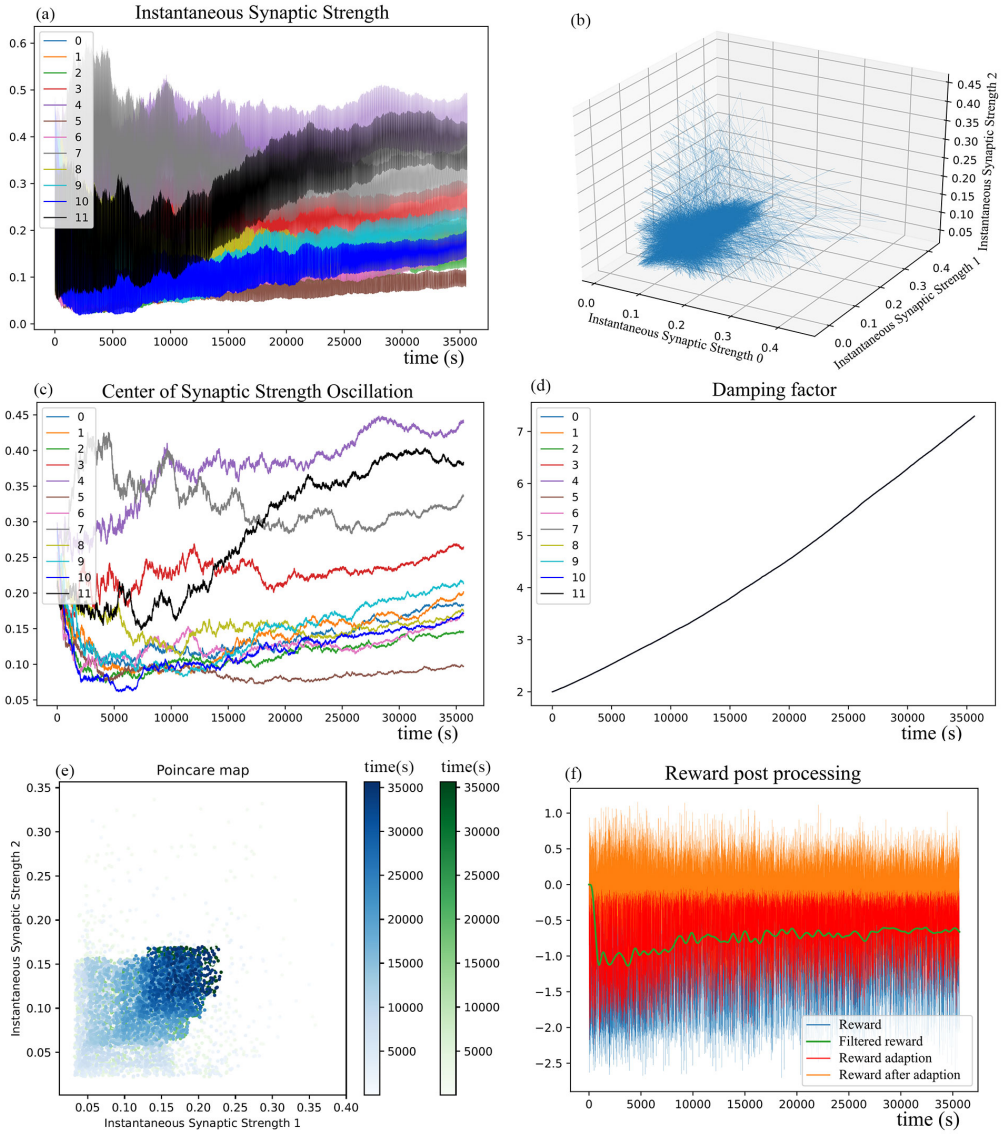


Figure 11: The simulation results of Dynamic Synapse in PuckWorld. The relationships between the labelled number of synapses and the sensor a synapse connects to are: 0,1: x-velocity; 2,3 y-velocity; 4,5 d_r in x; 6,7 d_r in y; 8,9 d_e in x; 10,11 $d_e$in y; in each case odd numbers are the inputs in the positive direction as explained in the text. (a) Instantaneous synaptic strength of 12 synapses. (b) The trajectory of the first 3 synaptic weights; the explored range gradually converges. (c) The centres of synaptic strength oscillations; (d) The damping factors of instantaneous synaptic strength oscillation. All lines overlap. (e) A Poincaré map of Dynamic Synapse. It is a section of (b) when instantaneous synaptic strength passes its centre of oscillation. Each point is an intersection of the trajectory and the plane defined by the centre of oscillation. The blue and green points show the intersections from two different directions. The intensity of colour indicates the time of intersections. (f) shows the reward R, adaption current I_{adapt} and Reward after adaption.

In addition, Fig 11 (b) shows the exploration of 3 instantaneous synaptic strengths. Fig 11(d) shows the damping factor of the oscillation of the instantaneous synaptic strengths. Fig 11 (e) is a Poincare map of the Dynamic synapse, i.e. the section of (b) when the instantaneous synaptic strengths 0 passed the centre of synaptic strength oscillation. It shows that the exploration is chaotic and unpredictable, and that the region of sampling shrinks during learning and the density of sampling increases during learning. (f) The line labelled Reward is the value R returned by the simulation enviroment by the reward function; The line labelled Filtered Reward is the low-pass-filtered R which shows the overall trend; the line labelled Reward Adaption is the adaption current I_{adapt}; the line labelled Reward after Adaption is the value of $k_R R - I_{adapt}$, which determines the modulator release and is more sensitive to variations of the reward than to the absolute value of the reward.

The source code for simulations of the model and experiments is available online `https://github.com/InsectRobotics/DynamicSynapsePublic`.

3. Discussion

We have proposed a model of operant learning based on continuous unpredictable synaptic strength fluctuations, with dynamics that are altered in response to a reinforcement signal. We illustrate the application of this principle to optimise the output, for given inputs, first in a simple linear neuron model, then to tune a recurrent CPG network to a target period, and finally to enable a spiking neural circuit embedded in an agent to improve performance in a continuous environment with dynamic reward and punishment.

An important property of our approach is that the source of variation that supports operant learning is continuous, unlike reinforcement learning algorithms that are based on random number generators, which have either discrete random outputs, or are partially predictable because of interpolation. By defining a system that has chaotic dynamics we can generate continuous motion without interpolation, so the unpredictability is continuous on any scale. An additional advantage over alternative synapse-level models for operant learning, such as the Hedonistic Synapse (Seung, 2003), are that the applications are not limited to a specific type of neural circuit or neural network. We have shown we can use our Dynamic synapse in both spiking and firing rate neural circuits, and the method can also be suitable for general online parameter optimisation, as it acts to scale the synaptic strength value

to the suitable ranges. It can also be applied to discrete systems by adjusting the time step to an appropriate range or by sampling. We plan to further explore the application of this model to a range of problems in robot learning and reinforcement learning.

A key difference between our model and previous models is that our model learns in parameter space but not action space. Previous models usually alter the synaptic strength based on the pattern of synapse activities (i.e. those conveying signals that led to reward), but our model directly learns the synaptic strengths that led to reward. As the synapse dynamics reflect recent states of the synapse, exploring parameter space enables our model to solve the credit assignment problem without an eligibility trace, which is necessary for some previous models, such as extended STDP models by Izhikevich (2007); Gurney et al. (2015). As the time scale of synaptic strength fluctuations is longer than synapse activity dynamics, the model can function with temporally distant reward. Exploring parameter space means that the learning concerns the overall function instead of the specific outputs of the neural circuits, so our model allows remodelling of synaptic connections independently from action potentials of neurons, which is a potentially powerful tool for neural computation.

We have proposed a possible grounding for the chaotic dynamics in the phenomena of receptor movement in dendritic spines. The model is inspired by recent evidence concerning the extent and mechanisms of these dynamics, but abstracted from the level of individual proteins to the level of the receptor flows between a dendrite and synapses as an integrated system. By focussing on postsynaptic receptor dynamics, our model can be related to synaptic mechanisms of short and long-term potentiation and depression (STP/LTP, STD/LTD). For example, the relations between STP and LTP as well as STD and LTD are similar to the relation in our model between the instantaneous synaptic strength and the centre of synaptic strength oscillation. The model can be expanded to explicitly explain some phenomena during STP, LTP, STD or LTD. For example, in STP-LTP model proposed in Xie et al. (1997), AMPA receptors are attracted toward the activated NMDA receptors when neurotransmitter is released, then a proportion of AMPA receptors diffuse again. This learning rule can be implemented by adding $k_{w1}n_T$ into the function describing the change of the amount of receptors in a synapse:

$$\dot{w}_i = \begin{cases} (v_i + k_{w1} n_T)\, c_d & \text{if } v_i > 0 \\ (v_i + k_{w1} n_T)\, \dfrac{w_i}{V_i} & \text{if } v_i < 0 \end{cases} \tag{11}$$

Where n_T is amount of the synaptic transmitter, k_{w1} is a coefficient. In this extended model, when neurotransmitter is released, the instantaneous synaptic strength (the number of receptors) will tend to increase, resulting in STP. When the instantaneous synaptic strength is higher than the centre of the oscillation, if modulator is released, the capacity of the synapse to contain receptors will increase. Because of the oscillation of the amount of receptors in the synapse, some of the receptors diffuse again. Because the capacity is increased, more receptors are held in the synapse, resulting in LTP.

The model in this paper represents postsynaptic dynamics in a simplified form, at the statistical level of receptor trafficking, allowing it to emulate some features of receptor flow dynamics and synapse dynamics. Modelling individual receptors is out of the scope of this study because it would not be relevant at the level of learning. However, the mathematical functions for the receptor dynamics in our model are not exclusive. As long as the receptor dynamics has the features of chaotic oscillation, and the centre of oscillation is controllable by our learning rule, our learning rule could work for alternative formulations. The model could be be extended to include more detail. For example, the receptor trafficking within the dendrite is assumed to be fast enough (compared to dendrite to synapse trafficking) to ignore its time constant. In reality, variations of AMPA receptor numbers on neighbouring dendrite spines are usually in the same direction (Zhang et al., 2015). This phenomenon could be modelled by taking account of the speed of receptor trafficking in the dendrite, which would have the consequence that neighbouring synapses would tend to have a similar concentration of receptors in the dendrite. Hence the receptor oscillation in neighbouring synapses would have a higher probability to be in similar phases than in distant synapses. Our model depends on several hypothetical assumptions, such as the form of the dynamics of receptor trafficking, dynamics of capacity to contain receptors, and the equilibrium point of receptor oscillation, which are not yet directly supportable from biological research. To understand the dynamics of receptor trafficking requires continuous observation of the collective motion of receptors and concentration change of receptors in dendrites and synapses

on timescales from seconds to hours. Similarly, understanding the dynamics of capacity to contain receptors requires continuous observation of actin flow between synapses and dendrites, size change of synapses and size change of postsynaptic density on similar timescales. Both types of observations are difficult but becoming experimentally more plausible, e.g. approaches of video microscopy in Zhang et al. (2015) and Esteves da Silva et al. (2015) continuously recorded the motions of proteins that can be observed as a group enabling the concentrations and flows to be understood. Observation of the phase relations between the oscillation of the receptors or structural components would be helpful for validating our model. In our model, we assume that the instantaneous weight leads the change of equilibrium point of receptor oscillation when the modulator is present. This could be tested by transplanting receptors to or from a synapse and giving modulator treatment, then observing if the synapse size or postsynaptic density changes. Thus several predictions arise from our model which we hope may be tested in future experiments.

However, the key concept presented here is not crucially dependent on the details of receptor trafficking. Other models of chaotic neurons or neural circuits suggest chaos exists in the membrane potential, and alternative chaotic processes in an animal could also possibly contribute to the generation of actions and learning with the same desirable properties of continuous unpredictability. Rather, the key properties are that the learning mechanism is entirely local to the synapse, and does not require an explicit tag for the Hebbian correlation of pre- and post-synaptic activity but rather allows this property to emerge from the behavioural or output consequences caused by the recent state of the circuit. That is, synapses that contribute to obtaining reward are strengthened; but this does not depend on the firing of either the pre- or post-synaptic neuron, except insofar as this is necessary to cause behavioural outputs that result in reward.

It is nevertheless interesting to consider a simple variation on the learning rule we have used to make synapses with active presynaptic neurons (neurons that have released neurotransmitter, indicating they have fired) learn actively (c.f. Eqs. 1 and 24):

$$\dot{w}_{ci} = \mathrm{k}_{w2}\,(w_i - w_{ci})\,n_M n_T \tag{12}$$

$$\dot{b} = \mathrm{k}_b b n_M n_T \tag{13}$$

where n_T is amount of the synaptic transmitter. With n_T, variation of synaptic strength of a synapse is proportional to the presynaptic neuron activity,

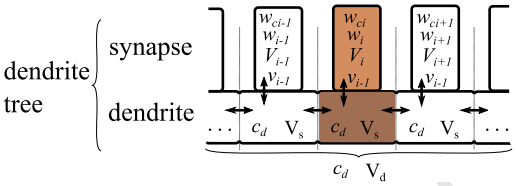


Figure 12: Schematic Diagram and Symbols of Dynamic Synapse. A schematic diagram of the dendrite tree; the main variables and parameters of the model are indicated. For the meaning of the symbols, see Table 1

which can help to improve the pertinence of learning to the inputs. For example, a neuron gets multiple inputs but only a small set of them is activated by a specific stimulus, and with this rule, the synaptic plasticity only applies between the neuron and these activated inputs. Note this is a 3-factor learning rule, depending on the correlation between the amount of the synaptic transmitter, the amount of modulator, and the difference between instantaneous synaptic strength and the centre of the oscillation. When the absolute value of the correlation is higher, the variation of the centre of the oscillation is more significant.

However, another possible learning rule could use the weighted average, rather than the product, of the synaptic transmitter and instantaneous synaptic strength:

$$\dot{w}_{ci} = \mathrm{k}_{w3}\left(q(\mathrm{k}_{w4} n_T - w_{ci} + \alpha) + (1-q)\left(w_i - w_{ci}\right)\right) n_M \tag{14}$$

where k_{w4} is a coefficent to fit the amount of transmitter to synaptic strength, q a proportion representing the relative weighting of these two factors, and α a constant. Notably, this rule can potentially account for Pavlovian classical conditioning, where the stimulus and reinforcer (neuromodulator) are presented together irrespective of the output. When $q = 1$, the learning rule is Pavlovian learning; when $q = 0$, the learning rule is operant learning. When q is close to 1, the learning process might look like classical conditioning with noise. Thus, classical and operant learning may coexist in the same neuron and even in the same synapse.

4. Methods

4.1. Overview

We first present a verbal description of how our model represents the alteration of synaptic strength in terms of the dynamic movement of receptors,

and then provide a precise mathematical formulation of the principle.

Two forms of receptor trafficking can move receptors between the synapses and the dendrite. Lateral diffusion creates a passive flow along a gradient from a high concentration region to lower concentration region. Endosomal trafficking acts as an active flow that can move receptors against the gradient. The active flow is formed by endosome transportation which carries numbers of receptors. Our model has a minimal form to capture the key phenomena. Endosomal trafficking is active transportation and is modelled with a positive feedback term which provides motive force, and two negative feedback terms which limit the speed of transportation. The negative feedbacks are the receptor concentration gradient, which is proportional to the concentration difference between a synapse and dendrite, and friction of endosome transportation, which is proportional to the endosome transportation speed. These properties together produce dynamic oscillation of the number of receptors in each synapse. Because of the concentration gradient, the equilibrium point of the dynamics of endosome transportation of a single synapse is when the concentration of receptor in the synapse is same as the concentration in the dendrite. It is also the equilibrium point of lateral diffusion. Note that because effects of receptor synthesis and degradation on receptor concentration are slower than receptor trafficking, they are assumed to have a negligible contribution to the dynamics. The proportion of receptors in endosomes is also ignored. Hence, in our model the total number of receptors in a dendritic tree is constant.

There are two factors in addition to receptor trafficking that could affect the concentration of receptors in each synapse: the size of the synapse and the number of receptors per unit area the synapse can accommodate. The size of the synapse is affected by the activity of actin. The number of receptors per unit area a synapse can accommodate is affected by scaffold-cytoskeleton complex. The two factors are not distinguished in the model but are jointly represented as the 'capacity' of the region to hold receptors. Thus, the equilibrium point of receptor motion can be altered by altering the capacity. The mechanism of learning in our model is to alter the capacity according to the following rule: whenever a neuromodulator signalling reinforcement is present, the instantaneous number of receptors in a synapse determines a change in its effective capacity, establishing a new equilibrium point nearer to that instantaneous value.

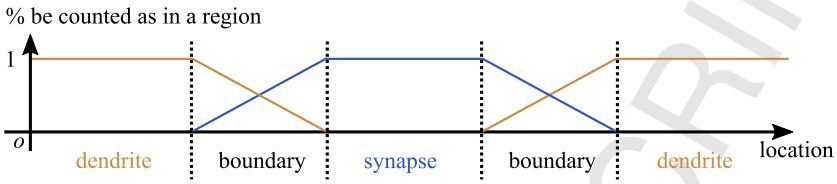


Figure 13: Justification for a continuous representation of the effects of receptor location between dendrite and synapse. The boundary between a synapse and dendrite can be considered wide and smooth, and as a receptor approaches the synapse, it can receive more neurotransmitters and contribute more to the synaptic strength. Rather than model the boundary area explicitly, we associate synaptic strength with the 'amount' of receptors a synapse contains, treated as a continuous variable.

4.2. Mathematical model

When the number of receptors per synapse is sufficiently large, their dynamics can be modelled statistically using differential equations (Holcman and Triller, 2006), e.g. like gas, which consists of free-moving molecules and uncertain intermolecular distance. However, even for a smaller number of receptors per synapse, we note their contribution to synaptic strength can be proportional to their distance from the centre of the synaptic cleft, due to diffusion of neurotransmitter (Fig 13). Thus, rather than explicitly represent discrete receptors and their positions, we represent the number of receptors in a synapse that currently contribute to its synaptic strength as a continuous 'amount'.

In the following equations, constants are represented by normal font and variables by italics (except v for membrane potential of integrate-and-fire neurons). The meanings of the symbols are shown in Table 1. The unit of time is millisecond.

The model assumes that the capacity of the dendrite to contain receptors is proportional to the number of synapses:

$$\mathrm{V_d} = \mathrm{N}\mathrm{V_s} \tag{15}$$

Where $\mathrm{V_d}$ is the capacity of a dendrite, N the number of synapses, and $\mathrm{V_s}$ a constant factor, which is the average capacity of a dendrite per synapse.

The concentration of receptors in the dendrite, c_d, is given by:

$$C_d = \mathrm{W_{total}} - \sum_{i=1}^{n} w_i/\mathrm{V_d} \tag{16}$$

where w_{total} is the (fixed) total amount of receptors in the dendrite tree; w_i is the amount of the receptors in the ith synapse; and V_d is the capacity of the dendrite.

We model the continuous flow of receptors between synapses and dendrite as a movement rate times the concentration of receptors on the source side:

$$\dot{w_i} = \begin{cases} v_i c_d & \text{if } v_i > 0 \\ v_i \dfrac{w_i}{V_i} & \text{if } v_i < 0 \end{cases} \tag{17}$$

where w_i is the amount of receptors of the ith synapse, w_i/V_i is concentration of receptors of the ith synapse, c_d the concentration of receptors in the dendrite, and v_i is the bidirectional movement rate, which is affected by lateral diffusion, endosomal trafficking and friction as described in the overview:

$$\dot{v_i} = 1/r \left(c_d - w_i/V_i + \mathrm{a}\,\mathrm{sign}(Vi) \times \sqrt[2]{|V_i|} - bv_i \right) \tag{18}$$

where v_i is bidirectional movement rate from dendrite to synapse (the direction from dendrite to synapse is positive); r is movement rate inertia , which represents factors (e.g. properties of actin) that drive receptors to keep their direction of flow; V_i is the capacity of ith synapse, which is affected by w_{ci}; $c_d - w_i/V_i$ is a term that represents the concentration difference between synapse and dendrite, which causes motion of receptors by diffusion; $\mathrm{a}\,\mathrm{sign}(Vi) \times \sqrt[2]{|V_i|}$ is positive feedback term of the movement, with positive feedback coefficient a; $-bv_i$ is a damping term with represents friction during the motion, with damping factor b.

As shown in Fig 12, the receptors also move between neighbouring dendrite regions by diffusion:

$$\dot{c}_{d_i} = q_d \left(c_{d_{i-1}} + c_{d_{i+1}} - 2c_{d_{i+1}} \right) \tag{19}$$

where q_d is a coefficient from concentration difference to concentration variation rate. In practice, we found that when the number of synapses is less than 33, modelling this this diffusive process has little effect. Hence, in the simulations in this paper, the diffusion is treated as instantaneous. For larger numbers of synapses, neglecting the dendritic diffusion can result in collapse of the chaotic dynamics, but these can be recovered if we run simulations with limited diffusion (results not included here).

As receptors diffuse in the dendrite tree, there is an equilibrium point when the concentration of receptors in a synapse and its neighbouring dendrite region are same. The equilibrium point forms the centre of synaptic strength oscillation, while the instantaneous synaptic strength oscillates around this point. We consider the effective strength of the synapse to be its equilibrium point, which can be established as follows. We assume that the receptors take a shorter time to diffuse between a synapse and its neighbouring region of the dendrite than to diffuse to regions in the neighbourhood of other synapses. Thus, in a short time interval, there is conservation of the amount of receptors in a synapse and its neighbourhood, and the equilibrium point is given by:

$$c_{ci}V_i/c_{ci}V_i + c_{ci}\mathrm{V_s} = w_{ci}/w_i + c_{d_i}\mathrm{V_s} \tag{20}$$

Where c_{ci} is the equilibrium concentration of receptors in ith synapse, w_{ci} is the equilibrium amount of receptors in ith synapse, w_i is the instantaneous amount of receptors in ith synapse, V_i is capacity of the ith synapse, c_{d_i} is concentration of the receptors in ith dendrite region and V_s is average dendrite capacity per synapse.

To set or alter the strength of a synapse, we alter w_{ci} . Solving the above equation for V_i, we get:

$$V_i = \mathrm{V_s}w_{ci}/c_d\mathrm{V_s} + w_i - w_{ci} \tag{21}$$

By updating V_i according to this function, the amount of receptors will converge to the given equilibrium value. Thus, we can define (or alter) the centre of synaptic strength oscillation. We can also alter the amplitude of oscillation around this centre by changing the damping factor b in equation 18.

These equations describe a system which contains multiple coupled second-order systems. A second-order system, such as a spring-mass-damper system, usually has the property of oscillation. When coupled together, they usually end in phase-locked oscillations, which means they have a fixed trajectory of oscillation. However, when the second-order systems include appropriate nonlinear functions, the system oscillates chaotically. In the model, the receptor trafficking between a synapse and dendrite is a second-order system. Multiple synapses are coupled by a dendrite, and updating of V_i is a nonlinear function. As we illustrate, the resulting oscillation appears to be chaotic. Because chaotic motion has a very complex, unpredictable and ergodic solution, the chaotic changes in synaptic strength can explore an output space for a neuron or neural circuit. Simulations are shown in the Results section.

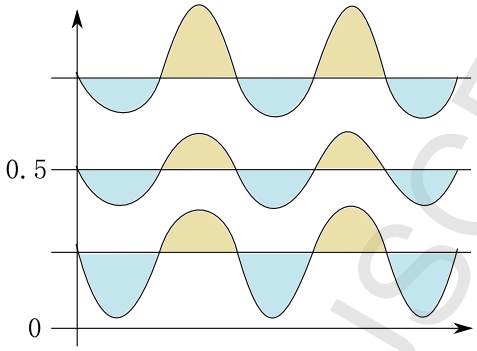


Figure 14: The bias of oscillation at different centre of oscillation. The curves are instantaneous synaptic strength, which oscillate around centres of synaptic strength oscillation (shown as straight lines).

As described in the Results section, a simple learning rule for this system is:

$$\dot{w}_{ci} = \mathrm{k}_w(w_i - w_{ci})n_{\mathrm{M}} \tag{22}$$

where n_M is amount of a neuromodulator that represents reward, and k_w is a coefficient controlling the learning rate. In practice we need to slightly modify this rule to compensate for a biased drift in synaptic strength. If, during an oscillation period, the integrated values of the differences between instantaneous synaptic strength and the centre of oscillation on each side is not equal (as shown in Fig 14, the sizes of adjacent yellow and blue coloured areas), uncorrelated modulator release (e.g. the release experienced by a synapse that is not making any useful contribution to satisfying the value function) can cause the centre of oscillation to become biased during long training times. During learning, if the centre of oscillation changes in a small range, the rate of bias can be approximated as a constant. To compensate it, a learning rule with compensation can be applied:

$$\dot{w}_{ci} = \begin{cases} \mathrm{k}_w\left(w_i - w_{ci}\right) n_M \left(1 + \mathrm{k}_{wc}\right) & \text{if } w_i > w_{ci} \\ \mathrm{k}_w\left(w_i - w_{ci}\right) n_M & \text{else} \end{cases} \tag{23}$$

where k_{wc} is a constant factor to compensate the bias. However, if the centre of oscillation changes in a larger range, the bias is variable, and cannot be compensated using the above rule. In our model, this bias is towards positive values for a centre of oscillation above 0.5, and negative values below 0.5. As a consequence there can be a positive feedback effect that accelerates learning.

To allow learning to converge, the learning rule should also reduce the oscillation amplitude. When the modulator is present, damping factors also increase:

$$\dot{b} = \mathrm{k}_b b n_M \tag{24}$$

where b is the damping factors, k_b a coefficient.

5. Acknowledgments

This work was supported by FP7 FET-Open project Minimal. We thank Matthieu Louis for discussions of earlier versions of this work.

References

Allam, S. L., Bouteiller, J. M. C., Hu, E. Y., Ambert, N., Greget, R., Bischoff, S., Baudry, M., Berger, T. W., 2015. Synaptic efficacy as a function of ionotropic receptor distribution: A computational study. PLoS ONE 10 (10), 1–20.

Allison, D. W., Gelfand, V. I., Spector, I., Craig, A. M., 1998. Role of Actin in Anchoring Postsynaptic Receptors in Cultured Hippocampal Neurons: Differential Attachment of NMDA versus AMPA Receptors. J. Neurosci. 18 (7), 2423–2436.
URL http://www.jneurosci.org/content/18/7/2423.short

Angulo-Garcia, D., Torcini, A., 2014. Stable chaos in fluctuation driven neural circuits. Chaos, Solitons and Fractals 69, 233–245.
URL http://dx.doi.org/10.1016/j.chaos.2014.10.009

Brembs, B., 2003. Operant conditioning in invertebrates. Current Opinion in Neurobiology 13 (6), 710–717.

Canavier, C. C., Clark, J. W., Byrne, J. H., 1990. Routes to chaos in a model of a bursting neuron. Biophysical journal 57 (6), 1245–51.
URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1280834{&}tool=pmcentrez{&}rendertype=abstract

Cash, D., Carew, T. J., 1989. A quantitative analysis of the development of the central nervous system in juvenileAplysia californica. Journal of Neurobiology 20 (1), 25–47.
URL http://doi.wiley.com/10.1002/neu.480200104

Cavalieri, L., Koçak, H., 1994. Chaos in biological systems. Journal Theorical Biology 169 (1985), 179–187.

Choquet, D., Triller, A., 2013. The dynamic synapse. Neuron 80 (3), 691–703.
URL http://dx.doi.org/10.1016/j.neuron.2013.10.013

Cingolani, L. a., Goda, Y., 2008. Actin in action: the interplay between the actin cytoskeleton and synaptic efficacy. Nature Reviews Neuroscience 9 (5), 344–356.
URL http://www.nature.com/doifinder/10.1038/nrn2373

Eckmann, J. P., Ruelle, D., 1985. Ergodic theory of chaos and strange attractors. Reviews of Modern Physics 57 (3), 617–656.

Esteves da Silva, M., Adrian, M., Schätzle, P., Lipka, J., Watanabe, T., Cho, S., Futai, K., Wierenga, C. J., Kapitein, L. C., Hoogenraad, C. C., 2015. Positioning of AMPA Receptor-Containing Endosomes Regulates Synapse Architecture. Cell Reports 13 (5), 933–943.

Frémaux, N., Sprekeler, H., Gerstner, W., 2010. Functional Requirements for Reward-Modulated Spike-Timing-Dependent Plasticity. Journal of Neuroscience 30 (40), 13326–13337.
URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.6249-09.2010

Gurney, K. N., Humphries, M. D., Redgrave, P., jan 2015. A New Framework for Cortico-Striatal Plasticity: Behavioural Theory Meets In Vitro Data at the Reinforcement-Action Interface. PLoS Biology 13 (1), e1002034.
URL http://dx.plos.org/10.1371/journal.pbio.1002034

Haselwandter, C. A., Calamai, M., Kardar, M., Triller, A., Azeredo Da Silveira, R., 2011. Formation and stability of synaptic receptor domains. Physical Review Letters 106 (23), 1–4.

Hausrat, T. J., Muhia, M., Gerrow, K., Thomas, P., Hirdes, W., Tsukita, S., Heisler, F. F., Herich, L., Dubroqua, S., Breiden, P., Feldon, J., Schwarz, J. R., Yee, B. K., Smart, T. G., Triller, A., Kneussel, M., 2015. Radixin regulates synaptic GABAA receptor density and is essential for reversal learning and short-term memory. Nature Communications 6, 6872.
URL http://www.nature.com/doifinder/10.1038/ncomms7872

Hayashi, H., Nakao, M., Hirakawa, K., 1983. Entrained, Harmonic, Quasiperiodic and Chaotic Responses of the Self-Sustained Oscillation of Nitella to Sinusoidal Stimulation. Journal of the Physical Society of Japan 52 (1), 344–351.

Holcman, D., Triller, A., 2006. Modeling Synaptic Dynamics Driven by Receptor Lateral Diffusion. Biophysical Journal 91 (7), 2405–2415. URL http://linkinghub.elsevier.com/retrieve/pii/S0006349506719567

Honkura, N., Matsuzaki, M., Noguchi, J., Ellis-Davies, G. C., Kasai, H., 2008. The Subspine Organization of Actin Fibers Regulates the Structure and Plasticity of Dendritic Spines. Neuron 57 (5), 719–729.

Ijspeert, A. J., 2008. Central pattern generators for locomotion control in animals and robots: A review. Neural Networks 21 (4), 642–653.

Inukai, H., Minami, M., Yanou, A., 2015. Generating chaos with neural-network-differential-equation for intelligent fish-catching robot. 2015 10th Asian Control Conference: Emerging Control Techniques for a Sustainable World, ASCC 2015.

Isaac, J. T., Nicoll, R. A., Malenka, R. C., 1995. Evidence for silent synapses: Implications for the expression of LTP. Neuron 15 (2), 427–434.

Izhikevich, E. M., 2007. Solving the distal reward problem through linkage of STDP and dopamine signaling. Cerebral Cortex 17 (10), 2443–2452. URL https://watermark.silverchair.com/bhl152.pdf?token=AQECAHi208BE49Ooan9kkhW{_}Ercy7Dm3ZL{_}9Cf3qfKAc485ysgAAAagwggGkBgkqhkiG9w0BB

Jaqaman, K., Kuwata, H., Touret, N., Collins, R., Trimble, W. S., Danuser, G., Grinstein, S., 2011. Cytoskeletal control of CD36 diffusion promotes its receptor and signaling function. Cell 146 (4), 593–606. URL http://dx.doi.org/10.1016/j.cell.2011.06.049

Jensen, G. D., 1963. Preference for bar pressing over "freeloading" as a function of number of rewarded presses. Journal of Experimental Psychology 65 (5), 451–454. URL http://content.apa.org/journals/xge/65/5/451

Kauer, J. A., Malenka, R. C., Nicoll, R. A., 1988. A persistent postsynaptic modification mediates long-term potentiation in the hippocampus. Neuron 1 (10), 911–917.

Koskinen, M., Hotulainen, P., 2014. Measuring F-actin properties in dendritic spines. Frontiers in Neuroanatomy 8 (August), 1–14.
URL http://journal.frontiersin.org/article/10.3389/fnana.2014.00074/abstract

Lau, C. G., Zukin, R. S., 2007. NMDA receptor trafficking in synaptic plasticity and neuropsychiatric disorders. Nature Reviews Neuroscience 8 (6), 413–426.

Mori, T., Nakamura, Y., Sato, M.-a., Ishii, S., 2004. Reinforcement Learning for a CPG-driven Biped Robot. Aaai 2004, 623–630.

Nargeot, R., Simmers, J., 2011. Neural mechanisms of operant conditioning and learning-induced behavioral plasticity in Aplysia. Cellular and Molecular Life Sciences 68 (5), 803–816.

Nobukawa, S., Nishimura, H., Yamanishi, T., Liu, J. Q., 2014. Analysis of routes to chaos in Izhikevich neuron model with resetting process. 2014 Joint 7th International Conference on Soft Computing and Intelligent Systems, SCIS 2014 and 15th International Symposium on Advanced Intelligent Systems, ISIS 2014, 813–818.

Petrini, E. M., Lu, J., Cognet, L., Lounis, B., Ehlers, M. D., Choquet, D., 2009. Endocytic trafficking and recycling maintain a pool of mobile surface AMPA receptors required for synaptic potentiation. Neuron 63 (1), 92–105.
URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2847611{&}tool=pmcentrez{&}rendertype=abstract

Roth, R. H., Zhang, Y., Huganir, R. L., 2017. Dynamic imaging of AMPA receptor trafficking in vitro and in vivo. Current Opinion in Neurobiology 45, 51–58.
URL http://dx.doi.org/10.1016/j.conb.2017.03.008

Sekimoto, K., Triller, A., 2009. Compatibility between itinerant synaptic receptors and stable postsynaptic structure. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics 79 (3), 1–13.

Sergé, A., Fourgeaud, L., Hémar, A., Choquet, D., 2003. Active surface transport of metabotropic glutamate receptors through binding to microtubules and actin flow. Journal of cell science 116 (Pt 24), 5015–5022.

Seung, S., 2003. Learning in Spiking Neural Networks by Reinforcement of Stochastics Transmission. Neuron 40, 1063–1073.
URL papers2://publication/uuid/5D6B29BF-1380-4D78-A152-AF8F233DE7F9

Sheng, M., Hoogenraad, C. C., 2007. The postsynaptic architecture of excitatory synapses: a more quantitative view. Annual review of biochemistry 76, 823–847.

Shepherd, J. D., Huganir, R. L., 2007. The Cell Biology of Synaptic Plasticity: AMPA Receptor Trafficking. Annual Review of Cell and Developmental Biology 23 (1), 613–643.
URL http://www.annualreviews.org/doi/10.1146/annurev.cellbio.23.090506.123516

Shouval, H. Z., Castellani, G. C., Blais, B. S., Yeung, L. C., Cooper, L. N., 2002. Converging evidence for a simplified biophysical model of synaptic plasticity. Biological Cybernetics 87 (5-6), 383–391.

Steingrube, S., Timme, M., Woergoetter, F., Manoonpong, P., 2011. Self-organized adaptation of a simple neural circuit enables complex robot behaviour. Nature Physics 6 (3), 16.
URL http://arxiv.org/abs/1105.1386

Storace, M., Linaro, D., De Lange, E., 2008. The Hindmarsh-Rose neuron model: Bifurcation analysis and piecewise-linear approximations. Chaos 18 (3), 1–11.

Sun, X., Milovanovic, M., Zhao, Y., Wolf, M. E., 2008. Acute and Chronic Dopamine Receptor Stimulation Modulates AMPA Receptor Trafficking in Nucleus Accumbens Neurons Cocultured with Prefrontal Cortex Neurons. Journal of Neuroscience 28 (16), 4216–4230.
URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0258-08.2008

Sussillo, D., 2014. Neural circuits as computational dynamical systems. Current Opinion in Neurobiology 25, 156–163.
URL http://dx.doi.org/10.1016/j.conb.2014.01.008

Tél, T., Gruiz, M., Kulacsy, K., 2006. Chaotic dynamics : an introduction based on classical mechanics. Cambridge University Press.

Triller, A., Choquet, D., 2005. Surface trafficking of receptors between synaptic and extrasynaptic membranes: And yet they do move! Trends in Neurosciences 28 (3), 133–139.

Wolf, R., Heisenberg, M., 1991. Basic organization of operant behavior as revealed in Drosophila flight orientation. Journal of Comparative Physiology A 169 (6), 699–705.

Xia, Z., Deng, H., Zhang, X., Weng, S., Gan, Y., Xiong, J., 2017. A central pattern generator approach to footstep transition for biped navigation. International Journal of Advanced Robotic Systems 14 (1), 1–9.

Xie, X., Liaw, J. S., Baudry, M., Berger, T. W., 1997. Novel expression mechanism for synaptic potentiation: alignment of presynaptic release site and postsynaptic receptor. Proceedings of the National Academy of Sciences of the United States of America 94 (June), 6983–6988.

Zhang, Y., Cudmore, R. H., Lin, D.-T., Linden, D. J., Huganir, R. L., 2015. Visualization of NMDA receptordependent AMPA receptor synaptic plasticity in vivo. Nature Neuroscience 18 (3).
URL http://www.nature.com/doifinder/10.1038/nn.3936

Table 1: Symbols in the equations .

Symbol	Explanation	Typical value
N	Number of synapses on a dendrite tree	an integer, > 3
$\mathrm{V_d}$	Capacity of a dendrite	$\mathrm{NV_s}$
$\mathrm{V_s}$	Average capacity of a dendrite per synapse	1
V_i	Capacity of the ith synapse	
$\mathrm{w_{total}}$	Total amount of receptors in the dendritic tree	
D_i	Occupation of a receptor in ith synapse	0 to 1
p	The constant coefficient for dimension conversion of the amount of receptors	
w_i	Instantaneous Synaptic strength of ith synapse	usually from 0.01 to 1
w_{ci}	Balance point of ith synapse	usually from 0.01 to 1
c_{d_i}	Concentration of the receptors in ith dendrite region	
$\frac{w_i}{V_i}$	Concentration of the receptors of the ith synapse	
v_i	Bidirectional movement rate from dendrite to synapse	
r	Movement rate inertia	3.5×10^6 to 2.5×10^7
a	The positive feedback coefficient of movement rate	170 to 850
b	The damping factor of movement rate	14000 to 2.6×10^7
q_d	The coefficient from concentration difference between neighbouring dendrite regions to receptor diffusion flux	
n_{M}	Amount of the modulator	usually from 0 to 1.5
k_w	A coefficient of balance point update speed	usually from 0.0003 to 0.002
k_{wc}	A constant factor to compensate the bias	0.4
k_b	A coefficient of damping factor update speed	usually from $10^{-7} to 10^{-8}$