Input-modulation as an alternative to conventional learning strategies

Esin Yavuz and Thomas Nowotny

School of Engineering and Informatics, University of Sussex
Falmer, Brighton, BN1 9QJ, United Kingdom
{e.yavuz,t.nowotny}@sussex.ac.uk
http://www.sussex.ac.uk

Abstract. Animals use various strategies for learning stimulus-reward associations. Computational methods that mimic animal behaviour most commonly interpret learning as a high level phenomenon, in which the pairing of stimulus and reward leads to plastic changes in the final output layers where action selection takes place. Here, we present an alternative input-modulation strategy for forming simple stimulus-response associations based on reward. Our model is motivated by experimental evidence on modulation of early brain regions by reward signalling in the honeybee. The model can successfully discriminate dissimilar odours and generalise across similar odours, like bees do. In the most simplified connectionist description, the new input-modulation learning is shown to be asymptotically equivalent to the standard perceptron.

Keywords: Reinforcement learning, olfactory system, spiking neural network

1 Introduction

Reinforcement learning is a learning paradigm in which appropriate actions are associated to sensory input guided by an evaluative feedback signal [16]. In computational models, this feedback is often considered to lead to modifications of the synapses between the outputs of the sensory processing cascade and the pre-motor regions that give rise to behaviour.

There are a number of regions associated with stimulus-reward associations in the human brain, and many different pathways are involved. On the other hand, insects are also capable of performing quite complex tasks even though their brains have much fewer brain regions and less than a million neurons. Honeybees, for example, rely on stimulus-reward associations for foraging, which is essential for their survival. Because of their extraordinary capabilities of solving complex learning tasks and the small size of their brains, they are a good animal model for studying the neural correlates of reinforcement learning [12].

The main brain regions of the honeybee olfactory system are the antennae, the antennal lobe (AL) and the mushroom bodies (MB). The olfactory receptor neurons (ORNs) in the antennae respond to olfactory stimuli and make synapses

with projection neurons (PNs) and local interneurons (LNs) in the AL. PNs then rely this stimulus-related information to the MBs. Even though the MBs are considered to be the main regions of stimulus identification and learning, bees have been shown to be capable of performing acquisition, but not consolidation, of elementary stimulus-reward associations even when the MBs are ablated [11] or their spiking activity is suppressed [3]. Moreover, injecting octopamine, a neurotransmitter which is known to mediate reward signalling in the insect brain, into the AL just after presenting an odour has been shown to be sufficient for conditioning [7]. Other studies have found that associative learning induces changes in the spiking activity of the neurons in the AL [5, 2, 15], not only in the MBs. These experiments suggest that reinforcement learning can be induced and evokes changes in the very early stages of the olfactory system, the AL, and that the MBs are not essential for simple elemental associative learning tasks.

Based on these observations, we have developed a spiking neural network model of the early olfactory system of the honeybee that does not require MBs to learn simple associations for appetitive absolute conditioning. The model uses elements of an earlier olfaction model [14], and includes additional mechanisms for reward modulation in the AL. Stimulus-reward associations are stored in plastic ORN-PN connections governed by a three-factor learning rule.

2 Model

The network connectivity is shown in Fig. 1. The model has four main layers: ORNs, PNs and LNs in the AL, and detector neurons (DNs), presumably located in the lateral protocerebrum. We have included 450 ORNs, 150 PNs, 30 LNs and 2 DNs, modelling roughly 1/5, or 30 glomeruli, of the AL. This choice was guided by the availability of experimental data for 30 glomeruli that are located dorsally and hence easily accessible for imaging [14]. Each glomerulus has five PNs and one LN associated to it. We have interpreted this as the substrate for five potential behaviours, and modify only the ORN-PN synapses of one of the PNs in response to reward. Others may be modified by other signals, e.g. in response to punishment. LN-LN and LN-PN synapses are inhibitory, all the other synapses are excitatory.

Each ORN expresses only one receptor type and each ORN type projects to the same glomerulus. ORN responses are modelled as Poisson spike trains with input-dependent rate. The rates are calculated as a function of the identity and concentration of odour input, in a rate model of binding, unbinding, activation and inactivation of receptors [13]. Details can be found in a previous study [14].

PNs and LNs are modelled as Hodgkin-Huxley type conductance based neurons [17], tuned to reproduce the electrophysiological data from honeybees [10]. We only modelled the homogeneous LNs which provide all-to-all inhibition, and excluded the heterogeneous LNs which connect to only a subset of glomeruli.

The membrane potential V_i of neuron i is given by:

$$C\dot{V}_i = -I_{\mathrm{Na},i} - I_{\mathrm{K},i} - I_{\mathrm{L},i} - I_{\mathrm{DC},i} - I_{\mathrm{syn},i}, \tag{1}$$

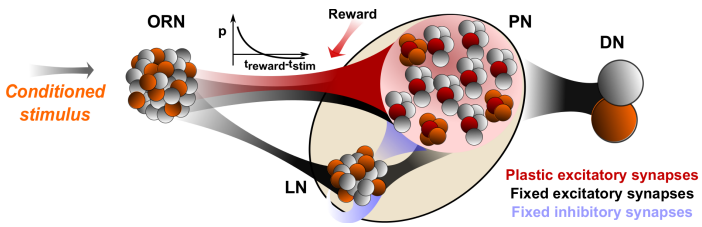

Fig. 1. Network connectivity. Presynaptic populations are at the light end of the synapses, and the postsynaptic populations are at the dark end. Each 5 PNs (and 1 LN) are members of the same glomeruli, and one of them makes reward-modulated plastic synapses with ORNs. LN-LN and LN-PN connections are inhibitory.

where C is the membrane capacitance, and $I_{\mathrm{DC},i}$ is a direct current injected into the neuron. The leak current is $I_{\mathrm{L},i} = g_{\mathrm{L}}(V_i - E_{\mathrm{L}})$ and the ionic currents $I_{\mathrm{Na},i}$, $I_{\mathrm{K},i}$ and $I_{\mathrm{M},i}$ are described by

$$I_{\mathrm{Na},i}(t) = g_{\mathrm{Na}} m_i(t)^3 h_i(t)(V_i(t) - E_{\mathrm{Na}}) \tag{2}$$

$$I_{\mathrm{K},i}(t) = g_{\mathrm{K}} n_i(t)^4 (V_i(t) - E_{\mathrm{K}}) \tag{3}$$

$$I_{\mathrm{M},i}(t) = g_{\mathrm{M}} z_i(t)^4 (V_i(t) - E_{\mathrm{K}}). \tag{4}$$

The synaptic current $I_{\mathrm{syn},i}$ received by neuron i is given by

$$I_{\mathrm{syn},i}(t) = (V_{\mathrm{syn}} - V_i(t)) \sum_j g_{\mathrm{syn},ij}(t) S_{ij}(t) \tag{5}$$

with a reversal potential of $V_{\mathrm{syn}} = 0$ mV for excitatory and -80 mV for inhibitory synapses. The synaptic activation variable S_{ij} is governed by

$$\dot{S}_{ij} = -\frac{S_{ij}}{\tau_{\mathrm{syn},ij}} + \sum_k \delta(t - t_j^{(k)}), \tag{6}$$

where $t_j^{(k)}$ is the time stamp of the kth spike of the presynaptic neuron j. Each activation and inactivation variable $m_i(t), h_i(t), n_i(t), z_i(t)$ satisfied first-order kinetics exactly as described in [14], equations (10) and (11).

The plastic synapses between ORNs and learning PNs were updated by a 3-factor learning rule. The "eligibility trace" p_{ij} is updated according to $p_{ij} \mapsto p_{ij} + F_{\mathrm{STDP}}(\Delta t_{\mathrm{spike}})$ whenever a pre- or post-synaptic spike occurs and otherwise decays exponentially according to $\dot{p}_{ij} = -(p_{ij} - p_{base})/\tau_p$. For simplicity, the STDP function was set to a constant, $F_{\mathrm{STDP}} = A$, if $-20 < \Delta t < 30$ and $= 0$ otherwise. The time window was chosen to match the STDP time window observed in locust Kenyon cells [1]. p_{ij} then drives changes in the synaptic conductance $g_{\mathrm{syn},ij}$ in conjunction with the reward signal R according to

$$\dot{g}_{\mathrm{syn},ij} = \frac{R \cdot p_{ij}}{\tau_{\mathrm{learn}}} - \frac{g_{\mathrm{syn},ij}}{\tau_{\mathrm{forget}}} \tag{7}$$

R is determined by the external experimental protocol in the form of a reward signal value $R_{\text{target}}(t)$ to which R approaches exponentially with a given time scale τ_{reward}:

$$\dot{R} = \frac{R_0 + R_{\text{target}} - R}{\tau_{\text{reward}}} \tag{8}$$

Here, R_0 is a negative baseline value for reward that causes responses in the absence of reward to lead to depression of $g_{\text{syn},ij}$ and hence extinction of previous memories. In a steady state without reward, both R and p_{ij} are negative; therefore the total effect on $g_{\text{syn},ij}$ is positive, resulting in recovery. The model behaviour is summarised in Table 1.

Table 1. Model behaviour as a function of reward (R) and eligibility (p)

	R <0	R >0
p <0	+ (recovery)	- (inactivation)
p >0	- (extinction)	+ (reinforcement)

The model was simulated on the GeNN GPU-accelerated modelling framework [18] [1].

3 Results

In order to test the performance of the model for discrimination and generalisation when forming associations between odours and reward, we tested responses to 2-Hexanol (2-Hex) against responses to 1-Hexanol (1-Hex) and 2-Octanol (2-Oct). According to behavioral [6] and calcium imaging [4] studies, 2-Hex and 1-Hex are similar and should lead to generalisation and 2-Oct is dissimilar and should be discriminated. When tested against 2-Hex conditioning, the behavioral generalisation probability of bees was 75.0% for 1-Hex and 37.5% for 2-Oct [6].

The absolute conditioning protocol used here consists of five consecutive presentations of an odour paired with a reward signal (A+). The odour is presented for 4 seconds, and the reward signal is introduced 2 seconds after the stimulus onset. The odour is presented for 3 seconds and the reward is presented for 3 seconds. After the five consecutive presentations of A+, a second odour is presented without any reward (B-). Following B-, the first odour is presented again three times without sugar pairing (A-).

As a result of conditioning, the glomeruli that are active during A+ increase their firing, while the glomeruli that are not active decrease their firing. The temporal evolution of the PN responses during the absolute conditioning protocol is shown in Fig. 2 for dissimilar odours, and in Fig. 3 for similar odours. Glomerulus 15 responds to 1-Hex but not to 2-Hex, therefore its synapses are weakened

[1] The code and the parameter values are available at `https://github.com/esinyavuz/Input-Modulation-Learning`.

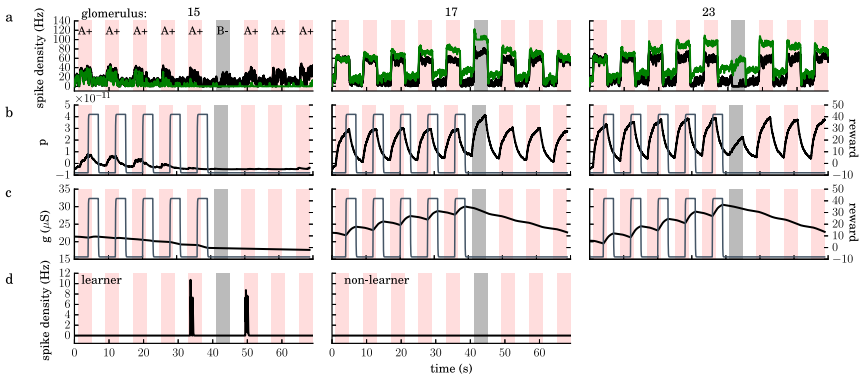

Fig. 2. Glomerular responses and coresponding learning parameters in time, during the absolute conditioning protocol for 2-Hex as the conditioning odour (A) and 2-Oct as the test odour (B), for three glomeruli. **a)** PN responses as spike density function of spike trains. Responses of the neuron that has learning synapses is shown in green, other neurons with simple synapses is shown in black. **b)** Eligibility traces and the reward signal. **c)** ORN-PN conductances **d)** DN responses.

during A+, which results in suppression of this glomerulus during B- (Fig. 3**a**, left). On the other hand, glomerulus 23 responds to 2-Hex and 1-Hex but not to 2-Oct, therefore learning results in a slight increase of its response to 2-Oct (Fig. 2**a**, right). Other glomeruli that have average response levels are slightly modulated, according to their eligibility that depends on the level of their activity (Fig. 2**b**). The resulting change in conductance is shown in Fig 2,3**c**. Changes in the spiking activity is then detected by the learner DN: it starts to fire after the 4th or the 5th conditioning trial (Fig. 2,3**d**), due to random initialisation of the conductances. During the test trial, it responds to the similar odour (Fig. 3**d**) but not to the dissimilar odour (Fig. 2**d**), which shows that the model could succesfully learn to discriminate the dissimilar odour while it generalises to the similar odour. After the first A- trial, the DN stops to respond as the odour is not associated with reward anymore, which is a phenomenon known as extinction.

4 Discussion

We presented a novel model for associative learning which involves modulation of AL activity by plasticity. This is much more akin to sensory learning than to usual associative learning models, like the classical perceptron. There are a number of alternative mechanisms that could underlie this type of learning. It could be based on recurrent network activity which provides a type of short-term memory and would facilitate the recurrent activation of PNs relevant to a rewarded stimulus. However, this is somewhat unlikely given that persistent spiking activity is not observed neither in calcium imaging, nor in electrophysiological recordings [4, 10]. Another alternative hypothesis would be changes in

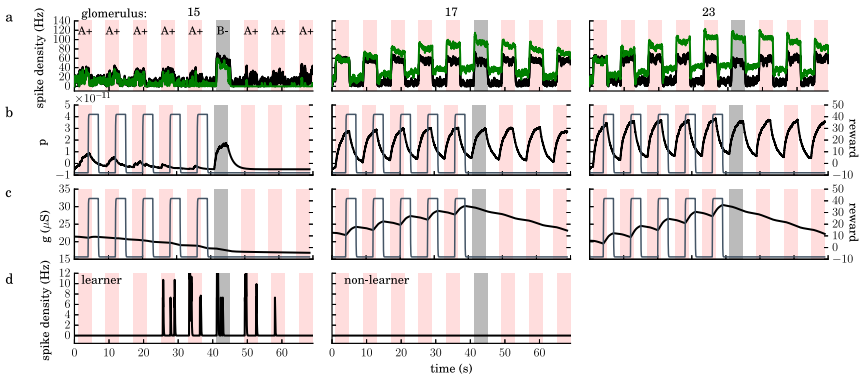

Fig. 3. Same as Fig. 2, for 2-Hex as odour A and 1-Hex as odour B.

neuronal properties of involved PNs or LNs, as has been observed in the snail feeding system [9]. The most likely substrate, however, are synaptic changes either between ORNs and PNs, as assumed here, or in the local network of the AL. This view is supported by the observation that associative learning induces changes in the spiking activity of the neurons in the AL [5, 2, 15].

The model of associative learning presented here is unusual compared to classical models such as the perceptron. In essence, this novel learning model is like a perceptron in which the input neurons learn to respond differently to rewarded inputs and so encode the knowledge of the world rather than modifying the synapses towards output neurons to achieve this. Figure 4 illustrates the essence of the two different solutions. It is natural to ask whether the two solutions are related and how they compare in classifying an input pattern against a backdrop of non-rewarded background patterns. To show this, we reduce the two models to a minimal connectionist description with binary variables as in [8].

For the perceptron, the responses are given by $y_i = \Theta\left(\sum_j w_{ij}x_j - \theta_{\mathrm{DN}}\right)$, where the input neurons x_j are the PNs, w_{ij} is a binary connection matrix, and θ_{DN} the firing threshold. The PNs governed by $x_j = \Theta\left(c_j \sum_k r_{jk} - \theta_{\mathrm{PN}}\right)$, and $r_{jk} \in \{0, 1\}$ are the responses of ORNs of type j, $k = 1, \ldots, N_{\mathrm{ORN}}$, and c_j is a synaptic connection strength from ORNs to PNs, the same for all k (and potentially also all j). θ_{PN} is the firing threshold of PNs. Learning takes place through changes in w_{ij}, e.g. by applying this simple stochastic binary learning rule [8] for a rewarded stimulus:

$$w_{ij}(t+1) = \begin{cases} 1 & \text{with probability } p_+ \text{ if } y_i = 1 \text{ and } x_j = 1 \\ 0 & \text{with probability } p_- \text{ if } y_i = 1 \text{ and } x_j = 0 \\ w_{ij}(t) & \text{otherwise} \end{cases} \tag{9}$$

It is straightforward to see that if the same pattern $\hat{\mathbf{x}} = (\hat{x}_j)$ is applied repeatedly and paired with an activation of y_i, then, eventually, the connectivity will equal $\hat{\mathbf{x}}$, i.e. $w_{ij} = \hat{x}_j$ for all j [8]. The separation of the target input $\mathbf{x}$ from other

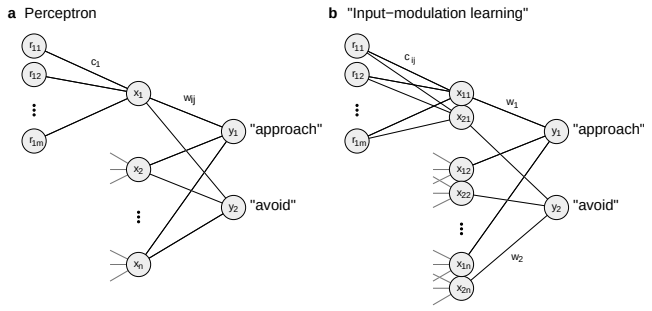

Fig. 4. Perceptron (**a**) compared to "input modulation learning" (**b**) introduced here. In the perceptron, the weights w_{ij} change during learning and there is only one copy of the "input pattern" in the PNs, and hence only one set of input neurons. In **b**, there are multiple copies of PNs, and learning takes place in the input synapses, c_{jk}.

inputs $\mathbf{x}$ then depends on the overlap of the other inputs with $\mathbf{x}$ and the value of θ_{DN}. In particular, the total input to y_i for an input pattern $\mathbf{x} = (x_j)$, is given by $\sum_j \hat{x}_j x_j$ or, equivalently, $\hat{\mathbf{x}} \cdot \mathbf{x}$.

On the other hand, for the learning system presented here, the responses are given by $y_i = \Theta\left(w_i \sum_j x_{ij} - \theta_{\mathrm{DN}}\right)$, the PN activitiy is $x_{ij} = \Theta\left(c_{ij} \sum_k r_{jk} - \theta_{\mathrm{PN}}\right)$, and it is the input conductances c_{ij} that change during learning. In the same minimalistic connectionist manner as above [8], an apropriate simplified description of the learning rule would be that, if reward is present,

$$c_{ij}(t{+}1) = \begin{cases} 1 & \text{with probability } p_+ \text{ if } r_{ij} = 1 \text{ and } i \text{ is the "reward pathway"} \\ 0 & \text{with probability } p_- \text{ if } r_{ij} = 0 \text{ and } i \text{ is the "reward pathway"} \\ c_{ij}(t) & \text{otherwise} \end{cases} \tag{10}$$

If no reward is present, all synaptic strengths c_{ij} remain unchanged. This scheme assumes, that there is one output neuron y_i per reward or punishment pathway, or, equivalently, each type of action, e.g. "approach" and "avoid". With the same argument as in [8] for the learning rule (9), the outcome of repeated i type reward for a single input pattern $\hat{\mathbf{r}}$ would be $c_{ij} = \hat{r}_j$ for all j. In such a case, we can assume that θ_{PN} is such that $r_{jk} = 1$ would lead to $x_{ij} = c_{ij}$, i.e. PNs x_{ij} spike if the corresponding receptors r_{jk} are active and $c_{ij} = 1$. If either the receptors are silent or $c_{ij} = 0$, no input will be received and no spiking will occur. The total input to an output neuron y_i hence depends essentially on the overlap of an input $\mathbf{r} = (r_j)$ with $\hat{\mathbf{r}}$, in particular, this input strength is $\sum_j \hat{r}_j r_j$ or the scalar product $\hat{\mathbf{r}} \cdot \mathbf{r}$, which is the same expression as above for the perceptron.

Our results suggest that input-modulation and the standard perceptron lead to the same results in their essentially reduced form, therefore the two approaches are equivalent. This indicates that learning can happen via different mechanisms than the traditionally studied ones. Investigating different strategies could provide insights to why the bees evolved to use this unusual mechanism, which is promising for development of novel algorithms for reinforcement learning.

Acknowledgments. This work is supported by the EPSRC (Green Brain Project, grant number EP/J019690/1) and Human Frontiers Science Program, grant number RGP0053/2015.

References

1. Cassenaer, S., Laurent, G.: Hebbian STDP in mushroom bodies facilitates the synchronous flow of olfactory information in locusts. Nature 448(7154), 709–713 (2007)
2. Denker, M., Finke, R., Schaupp, F., Grün, S., Menzel, R.: Neural correlates of odor learning in the honeybee antennal lobe. Eur J Neurosci 31(1), 119–133 (2010)
3. Devaud, J.M., Blunk, A., Podufall, J., Giurfa, M., Grünewald, B.: Using local anaesthetics to block neuronal activity and map specific learning tasks to the mushroom bodies of an insect brain. Eur J Neurosci 26(11), 3193–3206 (2007)
4. Ditzen, M.: Odor concentration and identity coding in the antennal lobe of the honeybee Apis Mellifera. Ph.D. thesis, Freie Universität Berlin (2005)
5. Faber, T., Joerges, J., Menzel, R.: Associative learning modifies neural representations of odors in the insect brain. Nat Neurosci 2(1), 74–78 (1999)
6. Guerrieri, F., Schubert, M., Sandoz, J.C., Giurfa, M.: Perceptual and neural olfactory similarity in honeybees. PLoS Biol 3(4), e60 (2005)
7. Hammer, M., Menzel, R.: Multiple sites of associative odor learning as revealed by local brain microinjections of octopamine in honeybees. Learn Memory 5(1), 146–156 (1998)
8. Huerta, R., Nowotny, T., Garcia-Sanchez, M., Abarbanel, H.D.I., Rabinovich, M.I.: Learning classification in the olfactory system of insects. Neural Comput 16, 1601–1640 (2004)
9. Kemenes, I., Straub, V.A., Nikitin, E.S., Staras, K., O'Shea, M., Kemenes, G., Benjamin, P.R.: Role of delayed nonsynaptic neuronal plasticity in long-term associative memory. Curr Biol 16(13), 1269 – 1279 (2006)
10. Krofczik, S., Menzel, R., Nawrot, M.P.: Rapid odor processing in the honeybee antennal lobe network. Front Comput Neurosci 2 (2008)
11. Malun, D., Giurfa, M., Galizia, C.G., Plath, N., Brandt, R., Gerber, B., Eisermann, B.: Hydroxyurea-induced partial mushroom body ablation does not affect acquisition and retention of olfactory differential conditioning in honeybees. J Neurobiol 53(3), 343–360 (2002)
12. Menzel, R.: The honeybee as a model for understanding the basis of cognition. Nat Rev Neurosci 13(11), 758–768 (2012)
13. Münch, D., Schmeichel, B., Silbering, A.F., Galizia, C.G.: Weaker ligands can dominate an odor blend due to syntopic interactions. Chem sens p. bjs138 (2013)
14. Nowotny, T., Stierle, J.S., Galizia, C.G., Szyszka, P.: Data-driven honeybee antennal lobe model suggests how stimulus-onset asynchrony can aid odour segregation. Brain Res 1536, 119–134 (2013)
15. Rath, L., Galizia, C.G., Szyszka, P.: Multiple memory traces after associative learning in the honey bee antennal lobe. Eur J Neurosci 34(2), 352–360 (2011)
16. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT press (1998)
17. Traub, R.D., Miles, R.: Neuronal networks of the hippocampus, vol. 777. Cambridge University Press (1991)
18. Yavuz, E., Turner, J., Nowotny, T.: GeNN: a code generation framework for accelerated brain simulations. Sci Rep 6, 18854 (2016)