2.3.1. Hedonism and delayed reward

As a precursor to neo-Hebbian RL, Seung (2003) proposed "hedonistic" synapses modeled as stochastic processes modulated by a global reward signal (inspired by the study of dopamine neuromodulation). The spiking IF neurons employed in this framework encapsulated the concept of synaptic weighting into a probabilistic synaptic spike transmission formulation wherein learning was a relation on global reward and the probability ($p_{j,i}$) of a pre-synaptic (unit j) action potential generating the release of some amount of neurotransmitter to the post-synaptic (unit i) membrane neuroreceptors across the synapse; for the purposes of the model, this is viewed as successful spiking.

$$p_{j,i} = \frac{1}{1 + e^{-w_{j,i} - c_j}} \tag{11}$$

Eq. (11) formulates the probability of a pre-synaptic action potential successfully affecting the post-synaptic neuron as a logistic sigmoid function (having range (0, 1)) on the learned weighting (w) of that connecting synapse and the calcium concentration within the pre-synaptic neuron. The variable c represents a simple (and interchangeable) model of calcium dynamics for the pre-synaptic unit, increasing by a parameter Δc at the moment of pre-synaptic spike generation (regardless of whether that spike event triggers release of neurotransmitter to the synapse) and exponentially decaying by $dc/dt = -c/\tau_c$ thereafter. Plasticity for the weight component under this framework (Eq. (12)) was modulated by the global dopaminergic reward signal ($R(t)$) and an eligibility trace (Eq. (13)) which decays following $d\lambda/dt = -\lambda/\tau_\lambda$. Parameter η functions as a learning rate to scale weight updates.

$$\frac{dw_{j,i}}{dt} = \eta R(t) \lambda_{j,i}(t) \tag{12}$$

$$\Delta\lambda_{j,i} = \begin{cases} (1 - p_{j,i}) \text{ if spike neurotransmitter release succeeds} \\ -p_{j,i} \text{ if release fails} \end{cases} \tag{13}$$

The additive update rule applied to $\lambda_{j,i}$ during pre-synaptic spiking incorporates some dynamics of the eligibility trace used in TD(λ) (Eq. (9)), increasing with the accumulation of recent relevant activity and decaying with temporal distance. Seung (2003) further conceptualized this framework as a greedy approximation of gradient ascent through the parameter space of w, deviating from the additive indicator function employed in the TD(λ) to restrict the eligibility trace to have zero mean so as to prevent bias in the weight traversal of that search space. A biologically plausible implication (from an operant conditioning perspective) of their formulation is the following consequence to plasticity with respect to rewards: recent and successful spike propagations relative to a positive reward signal result in LTP at the synapse and successful spike propagations followed by a negative reward induce LTD. Conversely, firing failures preceding a positive reward result in LTD while the same failures before negative rewards induce LTP. This is in contrast to the eligibility tracing of Eq. (9), which is strictly non-negative and would lead only to LTP given positive rewards and only to LTD under negative ones.

2.3.2. Distal rewards and credit assignment

Inspired by the formulations laid out in Seung (2003), Izhikevich (2007) incorporated dopaminergic reward directly into a modulated R-STDP learning strategy. The idea involved modulating the effects of STDP (LTP and LTD) on weight updates using a function of both direct environmental reward and the impact of those environmental reward signals on the changing concentration of dopamine over time (rather than a single, direct reward model as employed in Equation 12 by Seung (2003)).

This work introduced a more complex eligibility trace formulation, given in Eq. (14) where $STDP(\cdot)$ corresponds to an STDP plasticity rule like that given by Eq. (6). The eligibility trace is multiplied by the received temporally decaying reward signal $R(t)$, yielding Eq. (15) where $\frac{dR}{dt} = -\frac{R(t)}{\tau_{DA}} + \mathrm{DA}(t)$ and $\mathrm{DA}(t)$ is some function describing the dynamics of dopaminergic concentration. One example function modeling diffusive dopamine concentration dynamics used by Izhikevich (2007) was $\mathrm{DA}(t) = 0.5R(t - t_R)$ where t_R is the moment of receipt for the most recent reward value.

$$\frac{d\lambda_{j,i}}{dt} = -\frac{\lambda_{j,i}}{\tau_\lambda} + STDP(t_{post} \quad t_{pre})\delta(t \quad t^{(f)}) \tag{14}$$

$$\frac{dw_{j,i}}{dt} = \lambda_{j,i}(t) \cdot R(t) \tag{15}$$

Using their namesake Izhikevich spiking neuron model, Izhikevich (2007) employed the reward modulatory signal in conjunction with eligibility tracing to solve the distal reward credit assignment problem – the determination of assigning appropriate credit to synaptic weights with respect to their contribution towards rewarding or punitive performance over temporal scales – with spiking neurons in a framework consistent with STDP. The update rule in Eq. (14) can be viewed as an STDP-scaled form of the eligibility trace update from Eq. (9), which performs credit assignment for TD(λ) methods. In the context of neo-Hebbian RL, this trace performs credit assignment not for visited environmental states (as done by the TD(λ) method) but for activity over synaptic connections which precede rewards generated by the environment.

In this neo-Hebbian form, the Dirac delta function serves as the event indicator, comparable to $\mathbb{1}$ in Eq. (9), which increases the value for a given synaptic trace $\lambda_{j,i}$ at time $t = t^{(f)}$, where $t^{(f)}$ is the firing time of either the post-synaptic unit i (in the case of LTP) or the pre-synaptic unit j (for LTD), whichever occurs later in the spike pairing within the window for induction of STDP. This event-driven step-wise increase to the eligibility trace $\lambda_{j,i}$ is scaled by the value of $STDP(t_{post} \quad t_{pre})$, reflecting the dynamics of STDP with respect to the temporal difference between pre and post-synaptic spiking.

Many approaches to dopaminergic modulation of STDP since Izhikevich (2007) follow similar formulations, albeit adjusted to advance alternative aims beyond the credit assignment problem. For brevity in comparison of these spike-based neo-Hebbian works, we provide their formulations governing weight updates in Table 1.

Yusoffa and Grüning (2012) biased the effects of reward modulation on the eligibility trace using a learning rate α while training spiking units to associate delayed pairings of input stimuli, while Ozturk and Halliday (2016) achieved output spike train reconstruction by smoothing dopaminergic reward delivery with respect to average reward returns $\bar{R}$.

Potjans et al. (2011) formulated a more localized approximation of TD learning by modeling the reward signal $R(t)$ as fluctuations in dopamine concentrations relative to a baseline R_{base}, allowing the extension of an eligibility trace with an additional "activity" trace ε to reconstruct the TD error for each neural unit by interacting with the neuromodulatory dopamine concentration during weight updates.

Soltoggio and Steil (2013) showed that a rate-based equivalent of the R-STDP learning framework used in Izhikevich (2007) could achieve comparable results under a reward-modulated form of RCHP (see Section 2.1.1) in terms of learning in classical and operant conditioning tasks under delayed reward. Rather than using an explicit eligibility trace, as seen above in Eq. (14), Soltoggio and Steil (2013) deconstructed the synaptic weights to include long and short-term components such that $W_{j,i} = W^{lt}_{j\,i} + W^{st}_{j\,i}$ Changes to the short-term component of a given weight, $W^{st}_{j,i}$, which occur similarly to the updates of eligibility traces above, immediately impact the overall weighting of synaptic input at the post-synaptic neural unit but are not consolidated into the long-term weighting until the delivery of reward. This allows for the underlying Hebbian component of the three-factor neo-Hebbian formulation to perform unsupervised learning between rewarding events without inducing potentially erroneous permanent changes to the long-term weight.

$$\Delta W^{st}_{j,i}(t) = -\frac{W^{st}_{j\,i}(t)}{\tau^{st}} + RCHP_{j,i}(t) \tag{16}$$

Eq. (16) illustrates the dynamics of neo-Hebbian RCHP on the short-term weight component, where τ^{st} governs the decay rate of short-term plasticity changes and $RCHP_{j,i}$ corresponds to Eq. (3). Consolidation of the total weight value for each synapse occurs at the moment of reward delivery such that the long-term weight changes according to $\Delta W^{lt}_{j,i} = R(t)W^{st}_{j,i}$. While the modulatory signal $R(t)$ responsible for induction of short-term plasticity, in Soltoggio and Steil (2013) an essentially immediately impactful eligibility trace, was modeled as a discrete event, this does not preclude extension to continuous-time modeling akin to the dynamics of dopamine used by Izhikevich (2007). An upper-bound threshold for dopamine concentration could be used to induce long-term LTP, with a complimentary mechanism for LTD applicable in experiments which require it. The reward prediction error theory of dopamine which inspired much of computational RL theory is based largely on the study of dopamine transients, phasic activity by dopaminergic neurons which significantly deviate the extracellular concentration of the neuromodulator above or below its tonic baseline quantity in response to valued stimulation.

While not explicitly focused on the trade-off between exploration and exploitation in RL, Soltoggio and Steil (2013) did briefly consider the potential impacts of their synaptic weighting split. Repeated rare correlated activity at the synapse can allow for the short-term weights to grow rapidly without necessarily impacting the long-term component, as these changes to short-term plasticity decay rapidly. This may allow for the network to explore more extreme portions of the weight space during learning episodes in a temporary fashion, with repeated reward receipt inducing longer changes which encourage exploitative strategies.

Soltoggio (2015) extended the rate-based neo-Hebbian RCHP framework of Soltoggio and Steil (2013) with a focus on the issue of catastrophic forgetting in continual learning experiments. Their approach conceptualized the factoring of synaptic strength into short and long-term components as an approximate mechanism for hypothesis testing, using the modulatory signal $R(t)$ as evidence for or against the likelihood of a reward following stimulus-action pairs. Their newer formulation, termed Hypothesis Testing Plasticity (HTP), eschewed modeling LTD as a consequence of anticorrelated neural activity (the rate-based approximation of acausal STDP in RCHP) in favor of a consistent but weak form of weight depression provided by a slightly negative baseline value of dopamine – a strong contrast to the positive baseline value used in both Izhikevich (2007) and Soltoggio and Steil (2013). This negative baseline value for the modulatory signal continually induces LTD in the short-term weight components, which then require more consistent associations between experienced reward outcomes and stimulus-action pairs to grow large. We view the negativity of this baseline concentration of dopamine as a computationally expedient mechanism for replicating otherwise biologically plausible weak LTD in the absence of reinforcement by reward despite the clear implausibility of a negative baseline value for any neuromodulator.

When combined with a threshold for induction, the second major deviation of HTP from neo-Hebbian RCHP which solidifies short-term plasticity into the long-term weight component upon any reward delivery, this formulation protects the stability of the network parameters in the long-term weighting by only adopting permanent changes which have accumulated substantial evidence through trial-and-error.

$$\Delta w^{st}_{j,i}(t) = -\frac{w^{st}_{j,i}(t)}{\tau^{st}} + M(t)RCHP_{j,i}(t) \tag{17}$$

$$\Delta M(t) = -\frac{M(t)}{\tau^{M}} + \alpha R(t) - b \tag{18}$$

$$\Delta w^{lt}_{j,i}(t) = \beta \mathrm{H}\left(w^{st}_{j,i}(t) - \Phi\right) \tag{19}$$

Eqs. (17)–(19) illustrate the distinctions between neo-Hebbian RCHP and HTP. Short-term weights are continually updated by rare correlated activity following the RCHP rule as before, but are now also continually modulated by the function $M(t)$ which models the extracellular dopamine concentration as a decaying function of received rewards relative to a negative baseline value $-b$. Long-term plasticity is additionally modeled on a continual basis using the heaviside step function $\mathrm{H}(\cdot)$, which takes the value $+1$ for positive arguments and 0 elsewhere; the threshold for long-term LTP, Φ, ensures that positive argument values only occur when the short-term weight exceeds the minimum for induction. β is a consolidation hyperparameter similar to a learning rate that governs the speed of induction into long-term weight changes. The authors included this parameter to model temporal delays in biological plasticity changes, though they noted that instantaneous induction ($\beta = 1$) gave similar results. To model long-term LTD changes, a symmetric match for Eq. (19) is simple to produce using only negation and an appropriate lower bound (Soltoggio, 2015).

Acetylcholine is thought to play a role in a number of neural functions, including the consolidation of memories (Fink, Murphy, Zochowski, & Booth, 2013; Golden, Rossa, & Olayinka, 2016), spatial learning (Zannone, Brzosko, Paulsen, & Clopath, 2018), and attention to unexpected changes in stimulation (Brzosko, Zannone, Schultz, Clopath, & Paulsen, 2017). While R-STDP has been successfully employed in spiking models on tasks with stationary targets such as supervised classification (Hao, Huang, Dong, & Xu, 2020) or spike train sequence reproduction (Ozturk & Halliday, 2016), applications of R-STDP methods to RL problem domains with spiking neuron models have inherited some issues from their TD learning foundations. These relate to the reward landscapes of realistic environments, which are often sparse in terms of non-zero reward values (Machado et al., 2020) and dynamic (Hu et al., 2019).

Learning from extrinsic reward alone in environments with sparse and/or dynamic rewards has proven quite challenging for diverse sets of model agents. Intrinsic rewards have been introduced as a compensatory mechanism to aid learning when the reward space is insufficiently informative to guide exploitation (Gregor & Spalek, 2014; He & Zhong, 2018). While the application of intrinsic reward methods has largely been a feature of the gradient-based deep RL approach, we present in this section a brief overview of recent efforts to incorporate some form of intrinsic modulation of R-STDP with spiking neurons.

The majority of works addressing the concept of cholinergic modulation of R-STDP in SNNs employs the modeling of acetylcholine as a complementary factor to counterbalance the influence of dopamine modulation on STDP. Dopamine modulation which follows the general form outlined in the previous sections results in learning which closely follows TD methods. This entails a complete bias in weight updates towards exploitative strategies, as reinforcement alone only solves the credit assignment problem but does not directly encourage exploration of the state and action spaces in general (Sutton & Barto, 2017).

The formulation in Golden et al. (2016) modeled the purported dynamics of acetylcholine as dampening LTP by imposing a linearly decaying form of the learning rate parameter η (see Eq. (12) for a corresponding constant learning rate equation); as such, their plasticity mechanism (a standard STDP formula like Eq. (6)), eligibility trace (Eq. (23)), and consequential weight update rule (similar to Eq. (12) but with an STDP eligibility update rather than a probabilistic formula) did not differ in any substantive way from the dopaminergic formulations presented in Section 2.3.2.

$$\Delta\lambda_{j,i} = -\lambda_{j,i} + \eta STDP(t_{post} - t_{pre}) \tag{23}$$

Each training trial would incur a small decrement to η which simplistically modeled the effect of reduced levels of acetylcholine due to repeated stimulus exposure. This monotonic decrease in the learning rate was intended to capture the loss of agent surprise when returning to previously visited states due to trial repetition, with the decaying learning rate serving to enforce smaller weight updates as training progressed. The cause behind the findings in Golden et al. (2016), where a combination of dopaminergic and cholinergic modulation reduced convergence of performance in comparison to a dopamine reward baseline framework (where learning rate η remains constant), should be mathematically apparent.

We turn now to more advanced attempts at combining dopaminergic reward with cholinergic modulation by addressing the group of efforts made toward applying sequential neuromodulatory mechanisms (compared to the direct acetylcholine modulation of dopamine modulation embodied in the methods of Golden et al. (2016)). Brzosko et al. (2017), extending their previous work showing that dopamine signaling served to lengthen the time window dynamics under STDP, sought to encourage exploratory behavior by combining acetylcholine with reward signaling in simulations of dynamic environments. This sequential approach employed an alternating (see Eq. (25)) formulation of the effects of neuromodulation, with acetylcholine driving LTD on active synapses over timescales with low dopaminergic reward and with dopamine inducing LTP over eligible timescales, including those corresponding to periods of high cholinergic concentrations, as consistent with previous neuronal studies.

$$\Delta w_{j,i} = \eta A\Big(\sum_{t^{(f)}_{pre/post}} STDP(t_{post} - t_{pre}) \cdot \lambda_{j,i}\Big) \tag{24}$$

$$\Delta A = \begin{cases} -1 \text{ for } DA^-, ACh^+ \\ 1 \text{ for } DA^+, ACh^+ or ACh^- \end{cases} \tag{25}$$

The framework provided by Brzosko et al. (2017) improved upon the form of acetylcholine modeling employed by Golden et al. (2016) by applying an alternating rather than monotonically decaying learning rate η, where $\eta = 0.002$ in the presence of acetylcholine without dopamine and $\eta = 0.01$ during dopaminergic signaling. Further, their equation for the temporal decay of the eligibility trace λ alternated in effect according to the presence of dopamine, capturing the purported dynamics of dopaminergic stimulation on the STDP time window by following a longer exponential decay in the presence of dopamine (DA^+) and a typical exponential decay in its absence.

In their simulations requiring the learning agent to move to a locale associated with non-stationary reward, the addition of cholinergic modulation allowed the network to rapidly unlearn the previously memorized goal locations. In contrast, the dopamine-only baseline model frequently returned to formerly learned locations of reward long after the simulation had moved their position. This is consistent with the association between acetylcholine and exploratory behaviors and the reinforcement of reward coupled with dopamine that inspired their sequential neuromodulation framework.