2. Background

To enable a more rigorous treatment of the topics of neuromodulation and RL under neo-Hebbian plasticity, this section briefly introduces the standard notation for the computational models discussed in later sections, provides pertinent background detail on some concepts of significance to the topic of learning, and introduces the prevailing paradigms of artificial learning algorithms.

In the following sections, when discussing the dynamics for a given pair of neurons connected via synapse, we assign the pre-synaptic unit the index j and the post-synaptic neuron the index i. We may represent the output or "activation" of each unit by y_j and y_i, respectively. These activation values, which may be either continuous or discrete, will typically refer to the output of some (potentially non-linear) weighted function of the unit's inputs. The weights used for the calculation of a neuron's activation are the target of the learning or plasticity rules discussed herein, and we represent the general efficacy of the pre-synaptic unit j in affecting the activity of the post-synaptic unit i by the weight $w_{j,i}$. For a thorough introduction to the modeling of neuronal dynamics, including the wide variety of neuron models found in the texts we discuss later, we refer interested readers to the standard text found in Gerstner, Kistler, Naud, and Paninski (2014).

In the remainder of this review, we refer to computational models of neurons as being either spike-based or rate-based. This distinction is most easily understood from the perspective of the temporal granularity used in the modeling of neural activity. When we refer to spike-based neural network models, this signifies a finer temporal granularity of the neuron model that captures the timing $t^{(f)}$ of fired action potentials. This granularity contributes an increased biological realism to SNN simulations, capturing the temporal relationship whereby spikes from pre-synaptic neurons drive changes to the membrane potential of post-synaptic units that then may or may not cause the post-synaptic unit to fire at a later time.

Maintaining a record of the timing of spikes for both pre- and post-synaptic neurons allows for learning rules that consider both the temporal distance and order of spike events as factors, which we discuss in greater detail in Section 2.1.3. Although the focus of this review is limited to neo-Hebbian learning rules and the mechanics by which they manage the exploration–exploitation balance in RL tasks, there are a several prominent spike-based neuron models that underpin implementations of these neo-Hebbian networks. These include frequently used models in the computational neuroscience literature such as Integrate-and-Fire (IF) models (Lapique, 1907; Tuckwell, 1988), which are relatively expedient computationally and have been well studied in a number of model variations (Fourcaud-Trocmé, Hansel, van Vreeswijk, & Brunel, 2003; Hansel & Mato, 2001; Latham, Richmond, Nelson, & Nirenberg, 2000), as well as neuron models that more faithfully reproduce biological neural data such as the Izhikevich model (Izhikevich, 2003) and the Spike Response Model (SRM) (Gerstner, 1990; Gerstner, Ritz, & Van Hemmen, 1993). Paugam-Moisy and Bohte (2012) provide a highly accessible introductory treatment of these commonly employed neuron models.

Rate-based neurons model activity at a coarser temporal granularity than their spike-based counterparts, condensing the timing details of individual spikes into an average rate over uniform windows of time. Typically, simulations for rate-based ANNs do not actually model the spiking activity of biological neurons or the changes in membrane potential associated with it. Rather, rate-based neurons produce activation or output values as a (in most modern cases, non-linear Apicella, Donnarumma, Isgrò, and Prevete (2021)) function over a weighted sum of their pre-synaptic inputs. These activation values then represent the average spike count the unit would produce as output (to any post-synaptic neurons) over the next temporal window, although this aspect of biological realism is often neglected to allow for negative activation values that are understood to be inhibitory as negative rates over time are not biologically possible.

2.1. Adaptation and synaptic plasticity

Plasticity rules are attempts at computationally recreating the persistent changes in efficacy observed at synaptic junctions between pairs of neurons. These rules form the basis of learning algorithms for the neural network models in the context under review. As neo-Hebbian RL significantly extends Hebbian learning, in this section we provide an introductory treatment of the concepts and prominent formulations for this foundational class of bio-plausible learning algorithms.

2.1.1. Hebbian learning rules

Often cited and occasionally poorly paraphrased, the modern foundation of correlation-derived learning, Hebb's postulate (famously stated on page 62 of Hebb (1949)), proposes: "when an axon of cell A is near enough to excite cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased". The concept Hebb described is now referenced as long-term potentiation (LTP)

at the synaptic level, through which the relative efficacy of the connection between two neurons, typically treated as a weighting value of the connecting synapse, increases due to the effects of a pre-synaptic action potential contributing to the generation of a post-synaptic pulse.

We now know that correlated LTP alone is insufficient to capture the complex synaptic dynamics observed experimentally, so additional formulations for observed long-term depression (LTD) of the efficacy, or weight, of a synaptic connection are required (Dong, et al., 2012; Malenka & Bear, 2004). These may take the form of anti-Hebbian learning formulas, wherein pre-synaptic pulses driving post-synaptic spikes result in a net reduction of the impact of the synapse in question, or of frameworks that model the impact of improperly ordered spike pairs (a reversed spike ordering from standard Hebbian description). Other plausible mechanisms consider the concept of atrophy upon the synaptic connection as a factor that counteracts the monotone increases dictated by correlated LTP under Hebbian learning.

In the context of rate-based Hebbian learning, simultaneous correlated neural activation is typically (but not always) employed rather than an explicitly causal model which incorporates the timing of neural activations. If the rate of activation of two connected neurons are both coincidentally heightened/dampened, we may consider their activity to be correlated; if the activation rates diverge between the same pair of units, we consider their activity to be anticorrelated. These cases are typically associated as conditions to trigger LTP and LTD, respectively.

$$\tau_w \frac{dw_{j,i}}{dt} = f(w_{j,i})(y_j - y^{out})(y_i - y^{in}) \tag{1}$$

Eq. (1), adapted from Kuriscak et al. (2015), provides a stable form of rate-based Hebbian learning which models both LTP and LTD. Weight changes for the synapse connecting unit j to neuron i occur as a product of three expressions: (i) a function on the current weighting of the connection, which may introduce non-linear dynamics to weight changes; (ii) the difference between the current output firing rate of the post-synaptic neuron i and an upper bound threshold for output firing rates y^{out}; and (iii) the difference between the current input firing rate from pre-synaptic unit j and the respective upper bound threshold for input firing rates y^{in}. The use of differences between firing rates and threshold parameters, which can allow one or both of the latter multiplicative factors to take a negative sign, enables LTD to counteract unstable growth of the weight. This particular formulation implements both pre and post-synaptic gating when both threshold parameters are non-zero.

$$\Delta w_{j,i} = \alpha y_i (y_j - y_i w_{j,i}) \tag{2}$$

Oja's Hebbian learning rule, shown in Eq. (2), allows for a given neural unit to perform principle component analysis over its inputs (Oja, 1982) and is among the better studied variations on rate-based unsupervised Hebbian learning. The expression in parenthesis is considered the effective input to the neuron and serves to stabilize the growth of weights against divergence. Parameter α serves as a learning rate to scale the magnitude of weight updates.

While many other variations on the basic Hebbian learning rule exist, these rules operate on the same essential components – the current synaptic weight and the activation rates of both pre- and post-synaptic units. It is also possible to model causality – the effect of the pre-synaptic unit on the post-synaptic one – by using activation values which are not coincident in time. An example of this class of rate-based Hebbian learning is Rarely Correlating Hebbian Plasticity (RCHP), introduced in Soltoggio and Steil (2013). The original form of RCHP is expressed below in Eq. (3).

$$\Delta w_{j,i} = \begin{cases} +0.5 \text{ if } y_j(t - \Delta t) y_i(t) > \theta^+ \\ -1 \text{ if } y_j(t - \Delta t) y_i(t) < \theta^- \\ +0 \text{ otherwise} \end{cases} \tag{3}$$

This model of rate-based Hebbian learning induces specified, asymmetric weight adjustments for LTP and LTD when the product of post-synaptic activity (at the current time) with pre-synaptic input (over a small window of time Δt) exceeds (LTP) or fails to meet (LTD) established thresholds for their magnitude – θ^+ and θ^-, respectively. By ignoring common correlated activity between units – that which falls between the lower and upper thresholds specified – this formulation allows for RCHP to adjust the weighting only for near-coincidental neural activations that are highly likely to be correlated (or anticorrelated in the case of LTD). This rule can produce similar learning of weights to STDP in a highly efficient manner, as rate-based neurons are more efficient to simulate.

2.1.2. Differential hebbian learning rules

While the rate-based form of Hebbian learning described above implements synaptic plasticity changes on the basis of coincident pre- and post-synaptic activations, differential Hebbian learning (DHL) rules compute updates to synaptic weights using rates of change (derivatives) in neuron activations. This enables the learning rule to account not only for correlation but also for causation in the propagation of neural signals. We can assess this in Eq. (4), which is derived from the original formulation for this class of Hebbian learning in Kosko (1986).

$$\tau_w \Delta w_{j,i} = \dot{y}_i \dot{y}_j \tag{4}$$

In the equation above, $\dot{y}$ refers to the derivatives of the pre- (j) and post-synaptic (i) activities. This requires a differentiable model of neural activity that captures transient increases and decreases, which can be obtained for discrete event models by using an appropriate kernel on the activations. The use of such kernels, where necessary, ensures that a monotonic increase in the activation of a given unit temporally precedes a peak activation that is followed by a monotonic decrease. Using the overlap in these transients of neural activity enables a temporally symmetric modeling of LTP and LTD phenomenon. Considering the case when pre-synaptic neuron j begins to increase in activation (signifying that $\dot{y}_j > 0$) shortly before unit i does the same, $\dot{y}$ is positive for both pre- and post-synaptic units, yielding LTP of the synaptic connection during their overlapping transient increases. This is followed by a brief period of LTD when $\dot{y}_j$ becomes negative before $\dot{y}_i$ does. If we reverse the ordering of these transient increases, the cumulative change to synaptic weighting remains the same as both $\dot{y}_i$ and $\dot{y}_j$ are positive until, in this case, $\dot{y}_i$ becomes negative first. The net effect in both cases is LTP or increased weighting. Conversely, when either unit begins a transient increase in activity while the other is expressing a transient decrease, their overlap yields opposing signs and thus produces LTD.

To better reflect the causal relationship between pre- and post-synaptic activities, Porr and Wörgötter (2003) introduced a temporally asymmetric variant of DHL called isotropic sequence order (ISO) learning. This DHL variant models causality by replacing the derivative of the activation of the pre-synaptic unit $\dot{y}_j$ with the current activation, yielding Eq. (5).

$$\tau_w \Delta w_{j,i} = \dot{y}_i y_j \tag{5}$$

This formulation by Porr and Wörgötter (2003) captures the same correlations in synaptic activity while enforcing the temporal ordering required to infer causality in pre-synaptic activity

driving post-synaptic responses. If pre-synaptic neuron j is active while unit i is becoming more active ($\dot{y}_i > 0$), we can infer that the activity of j is at least partially driving the increase in activity expressed by i and the ISO rule produces LTP. When j is active (but not necessarily becoming more active) while unit i is becoming less active ($\dot{y}_i < 0$), we know that the activity of j is not driving this change in the activity of i and the rule induces LTD as a consequence. Similarly, dampened activations by unit j when i is experiencing a transient increase yields LTD while the same lack of activity by j results in LTP when i is becoming less active.

Zappacosta, Mannella, Mirolli, and Baldassarre (2018) constructed a framework to unify the variety of first-order DHL rules proposed in the literature, yielding general differential Hebbian learning or G-DHL. Their formulation considers eight components, divided evenly into differential (both factors being derivatives as in Eq. (4)) and mixed (derivative and non-derivative as in Eq. (5)) additive factors. The total of eight factors is derived by decomposing the derivatives of pre- and post-synaptic units into their positive and negative components. Each of these factors can be manipulated via hyperparameter to influence their inclusion/exclusion (non-zero or zero), their direction of influence (LTP or LTD based on sign), and their contribution to weight updates (their magnitude). The flexibility afforded by this generalization of DHL enables the reproduction of many experimentally observed neural phenomenon, as demonstrated in Zappacosta et al. (2018), and it may be employed for both rate and spike-coded neural models.

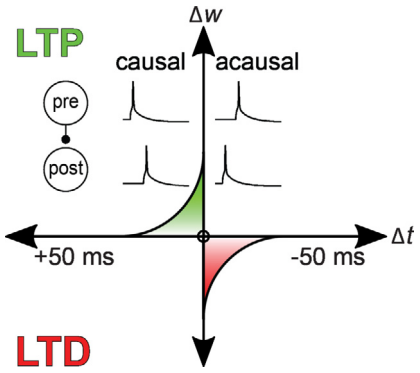

Fig. 1. Illustration of Δw (y-axis) with respect to Δt (x-axis) under a general STDP framework for both pre-then-post (causal, LTP, green) and post-then-pre synaptic pairings (acausal, LTD, red). Adapted from Markram, Gerstner, and Sjöström (2011). Note that the horizontal axis, corresponding to Δt, is reversed from the conventional left-to-right increase in value; the same is true in the originating source graph. As the magnitude of Δt increases between spike pairings, the corresponding synaptic changes to Δw become smaller — these spike pairings, pre-then-post on the left and post-then-pre on the right, are understood to be less indicative of a causal or acausal relationship between the spike pairing. Conversely, as Δt approaches 0 on the graph, indicating a shorter temporal interval between pre- and post-synaptic spikes, the STDP framework induces significantly stronger LTP (left, green) or LTD (right, red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.1.3. Spike-timing-dependent plasticity

STDP provides a framework for formulations of biological and artificial spiking neural systems consistent with the causal relationship central to Hebbian learning and compatible, by extension or modification, with anti-Hebbian phenomena. As the name implies, STDP rules employ the timings of spikes (in the simplest case pairs, as treated here) to determine the appropriate change in synaptic strength between the units involved. We present here the formulation for a basic pair-based STDP update rule. Consider a pair of neurons, j and i, connected as pre-synaptic and post-synaptic units, respectively. We denote the times of the events of their pre-synaptic and post-synaptic action potentials as t_{pre} and t_{post}, defining a measure on their temporal separation as $|\Delta t| = |t_{post} - t_{pre}|$. A simple update rule for the weight $w_{j,i}$, adapted from Gerstner et al. (2014), is expressed in Eq. (6).

$$\Delta w_{j,i} = \begin{cases} A_+(w_{j,i}) e^{\frac{-|\Delta t|}{\tau_+}} & \text{at } t = t_{post} \text{ for } \Delta t > 0 \\ A_-(w_{j,i}) e^{\frac{-|\Delta t|}{\tau_-}} & \text{at } t = t_{pre} \text{ for } \Delta t < 0 \end{cases} \tag{6}$$

This formulation permits both flexible Hebbian and anti-Hebbian dynamics through the selection of adaptation functions, $A_+(w_{j,i})$ and $A_-(w_{j,i})$, which may model alterations of efficacy as a function of the current synaptic weighting. The decaying exponential term, including the LTP and LTD decay constants τ_+ and τ_-, reflect the principle of temporal locality in STDP update rules; spike pairs relatively close in time (typically on the order of tens of milliseconds) are less coincidental and experimentally evoke stronger adaptive responses than more remote pairs.

This principle can be appreciated visually by assessing Fig. 1, which graphs the magnitude of plasticity changes with respect to the timing difference between pre and post-synaptic spiking for LTP and LTD. We refer to the distinction between LTP caused by a pre-before-post spike pair ordering and LTD arising from a post-before-pre spike pairing as causal and acausal forms of STDP, respectively. In the case of LTP under STDP, a pre-synaptic action potential contributing to the generation of a spike at the post-synaptic neuron shortly thereafter implies a causal relationship under Hebb's postulate. Pre-synaptic activity that follows a post-synaptic spike is inherently acausal, having not served to drive the post-synaptic pulse which preceded it. Markram et al. (2011) contain a highly accessible overview of the concepts and history behind the STDP framework that provides a deeper intuition on the biological considerations behind its development.

2.2. Learning paradigms

In regard to computational learning theory, there are three classic paradigms: unsupervised, supervised, and reinforcement. These broad theories on the nature of learning may also overlap in some frameworks. This section briefly treats on the distinctions of these approaches and their better-known methodologies.

2.2.1. Unsupervised

Having already introduced the concept of Hebbian learning, whereby connections are strengthened through correlated activity, the fundamental functional characteristic of unsupervised learning is both simple and yet foundational to more advanced schemes for learning. Unsupervised learning mechanisms such as Hebb's rule (Gerstner et al., 2014) for rate-encoded neural networks or STDP in spike-encoded variants operate on the principle of strengthening or weakening synaptic connections irrespective of consequent network activity. This type of learning does not incorporate any measure of correctness or utility, yet the identification of associations, whether simply correlational or indicative of causality, continues to serve as an underlying factor for more complex learning capabilities.

From an evolutionary perspective, the unsupervised aspect of neural learning logically precedes any capacity for feedback — the ability to distinguish similarities and differences in stimulation must be possessed prior to the development of stimulus–response dynamics, for example. Given this view of a natural progression in the development of learned reactions predicated on the precedence of unsupervised mechanics, we treat the ensuing learning paradigms as mechanisms implementing selectivity, among more intriguing dynamics, atop this foundation.

2.2.2. Supervised

Supervised learning methods incorporate the concept of correctness to induce selective responses; this requires an explicit error signaling mechanism in addition to a determined source of ground truth. The most prolific format of supervised training common in the literature is the backpropagation (BP) method (Rumelhart, Hinton, & Williams, 1986), famed for its deep learning successes in combination with the availability of large labeled datasets and efficient gradient of error calculations (Shrestha & Mahmood, 2019). Deep learning methods have achieved remarkable success across many domains of machine learning, including models for image classification with deep convolutional networks such as GoogLeNet (Szegedy, et al., 2015) and language translation with attention mechanics such as the Transformer architecture (Vaswani, et al., 2017). While other supervised learning approaches exist (Lee, Zhang, Fischer, & Bengio, 2015; Wang, Belatreche, Maguire, & McGinnity, 2014; Zenke & Ganguli, 2018), BP is the predominant form of supervised training for artificial learning algorithms and perhaps the most exemplary of the paradigm.

A more interesting and potentially pertinent variation on the concept of supervised learning is a self-supervised approach. While this approach has received more attention for its use in static generative modeling, such as in autoencoders (Hinton & Salakhutdinov, 2006) and generative adversarial networks (Goodfellow, et al., 2014), the use of these methods with recurrent networks can produce powerful sequential prediction models (Jawed, Grabocka, & Schmidt-Thieme, 2020). In these methods, the ground truth signal is not a handcrafted indicator of correctness (such as labels for classification tasks) but rather a withheld or otherwise unseen (to the learning agent) portion of the training data. For time series data, this requires the model to produce output intended to predict the next value, corresponding to some $t + \Delta t$, after having received inputs over the series up to the stimulation corresponding to the current time t. The predicted next stimulus may then be compared with the actual next input data from the series and many error correction-based learning algorithms, such as BP methods, may be used to improve the predictive prowess of the agent. This approach to learning over sequential data can be related to certain predictive or planning methods for RL, as seen in Pathak, Agrawal, Efros, and Darrell (2017) for example.

2.2.3. Reinforcement

As this work focuses on reinforcement learning (RL) with respect to bio-plausible learning models, we present here a brief introduction to the classic reinforcement learning formulations from the broader domain of machine learning, adapted from the modern text by Sutton and Barto (2017). In a neo-Hebbian context, RL builds upon the unguided coincidence detection of unsupervised learning by incorporating the concepts of reward and punishment, rather than the extension with evaluation by predetermined correctness employed by supervised learning methods.

The concept of reinforcement of learned behaviors is well established in the study of operant conditioning, whereby the voluntary response of an organism to a given stimulus is modulated either to increase or decrease the probability of that response in the future. This conditioning occurs through repeated observance of net positive (reinforcing, either through exposure to a pleasant stimulus or removal of an aversive one) or net negative (punitive, either through subjection to an aversive stimulus or elimination of a pleasant one) outcomes. In conjunction with early results from the study of dynamic programming, computational RL theory sought to develop learning agents capable of adapting to experiential feedback inherent to a defined environment, rather than through instruction by an explicit error signal. Accomplishing this form of learning requires an agent to both explore its environment and to learn to exploit the information gleaned from its interactions with the environment to most effectively maximize (minimize) the cumulative reward (punishment) to which it is subject over some typically variable temporal scale.

Temporal-difference (TD) methods are a central framework in the domain of RL. The simplest TD formulation, known as TD(0) for its one-step temporal window, calculates an update to the internal estimate of an environmental state's value $V(S_t)$ following observance of $V(S_{t+1})$ and any reward or punishment R_{t+1} generated due to the activity of the agent at the former state S_t; more abstractly, the underlying expectation of the value of the former state is updated by the experience obtained from interacting with that state through some available form of activity.

$$V_{t+1}(S_t) = V_t(S_t) + \alpha[R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)] \tag{7}$$

The one-step state-value update rule for TD(0) is given in Eq. (7), where α is a learning rate parameter, γ is a discounting factor accounting for the delay in obtaining the reward value of the successor state of the environment, and the term within the brackets is typically referenced as the TD error. These updates apply directly to the value function, an estimate of the actual long-term value of environmental states typically calculated by repeatedly sampling the environment through exploration of the state–action space. These value function updates impact the action selection policy through their role in determining the expected values of potential successor states. Note that an action with a potentially very high successor state value for a highly improbable state transition may not be selected by a greedy policy if a lower valued but more likely successor state is expected to offer higher cumulative rewards after taking a different action.

The one-step update of TD(0) can be generalized to account for a longer temporal window of experiential information on the value of a given environmental state by following the formulation of the TD(n) state value update rule, wherein the value of the state observed at time t is updated at time $t + n$ with a series of discounted returns generated through that n-step temporal window; the TD(n) state value update function is given in Eq. (8), noting that Eq. (7) corresponds to a reduction to $n = 1$.

$$\begin{aligned} V_{t+n}(S_t) = V_{t+n-1}(S_t) + \alpha[& \sum_{i=1}^{n} (\gamma^{i-1} R_{t+i}) \\ & + \gamma^n V_{t+n-1}(S_{t+n}) - V_{t+n-1}(S_t)] \end{aligned} \tag{8}$$

While the TD(n) method allows for effective value estimation using delayed rewards, the temporal duration of n is fixed. The TD(λ) method extends this approach by introducing an eligibility trace, denoted as λ, which allows for bootstrapping of rewards received in arbitrarily distant future states into the value estimate. The TD(λ) method is also more readily extensible to continuous-time modeling frameworks. Note that while Sutton and Barto (2017) use the symbol λ to represent multiple mathematical abstractions in various contexts throughout that text, we use λ exclusively to refer to an eligibility trace as is the standard in most recent works on neo-Hebbian RL.

Under the TD(λ) method, the eligibility trace λ for each state is used as a multiplicative factor on the TD error in the value update rule (Eq. (10)). The trace's value is a non-negative scalar which records how often a state has been visited and how recently these visits have occurred. The intuition behind eligibility tracing is that frequently visited states that precede rewards are more important for learning and their relevance has a temporal shelf life.

$$\lambda_t(s) = \gamma \lambda_t(s) + \boldsymbol{I}(S_t = s) \tag{9}$$

$$V_{t+1}(S_t) = V_t(S_t) + \alpha \lambda_t(S_t)[R_{t+1} + \gamma V_{t+1}(S_{t+1}) - V_t(S_t)] \tag{10}$$

The trace update in Eq. (9) captures this. The second term on the righthand side counts the number of times that a state has been visited using an indicator function $\boldsymbol{I}$ which takes the value 1 when the argument is true and 0 elsewhere. The first term causes the trace value to decay asymptotically to zero over time according to the trace decay parameter $\gamma \in [0, 1]$. As such, the value of λ for a given state reflects a function of both that state's visitation and its temporal relationship with delayed rewards, which is implicitly recorded by the amount of decay. Eq. (9) is adapted from Equation 7.5 in the first edition (Sutton & Barto, 1998) of the primary text used for this section, replacing a conditional equation with an equivalent indicator function. This particular formulation of eligibility tracing for computational RL was used as inspiration for synaptic eligibility tracing methods which can enable neo-Hebbian RL with distal rewards, as introduced in Section 2.3.2.

2.3. Modulation and reinforcement learning

We have given the relevant background for: (i) rate and spike-based neuron models, (ii) formulations for the basic Hebbian learning rules, and (iii) the main learning paradigms that can enable the extension of Hebbian learning rules. This section introduces a subset of the class of RL formulations commonly referred to as three-factor, or neo-Hebbian, learning rules. Neo-Hebbian mechanisms modulate (up or down) the change in strength between pre- and post-synaptic synapses, normally caused by a two-factor Hebbian rule, by incorporating the notion of value (or reward) as a third factor.

In its most basic form, neo-Hebbian RL alters standard Hebbian plasticity with various forms of scaling or gating in response to global reward signaling. The neuromodulator dopamine is often proposed to serve as this rapid signaling mechanism for reward in theories of behavioral learning in animals. This interpretation of dopaminergic neural activity has inspired a number of frameworks aiming to integrate the concept of TD reward errors with Hebbian learning theory (Frémaux & Gerstner, 2016; Gerstner et al., 2018). While alternative theories on the role of dopamine in learning have been proposed (discussed in later sections), a global reward signal is essential for extending Hebbian plasticity into more complex RL frameworks and its role must be understood before considering alternative modulating dynamics.

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 $\frac{dR}{dt} = -\frac{R(t)}{\tau_{\text{DA}}} + \text{DA}(t)$ and $\text{DA}(t)$ is some function describing the dynamics of dopaminergic concentration. One example function modeling diffusive dopamine concentration dynamics used by Izhikevich (2007) was $\text{DA}(t) = 0.5R(t - t_{\text{R}})$ where t_{R} is the moment of receipt for the most recent reward value.

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

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

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

Table 1
Summary of the formulations for R-STDP weight updates and eligibility trace calculations provided for comparison of the methods surveyed above. Note that $\Delta(t^{(f)}_{pre/post})$ corresponds to $t_{post} - t_{pre}$, the input term for spike pair-based STDP.

Source	Update rules
Potjans, Diesmann, and Morrison (2011)	$\Delta w_{j,i} = \alpha[\lambda_i \varepsilon_i][R(t) - R_{base}]$ $\Delta\lambda_i = -\frac{1}{\tau_\lambda}(\lambda_i - \sum_{t_i^{(f)}}(\delta(t - t_i^{(f)})))$ $\Delta\varepsilon_i = \frac{\varepsilon_i}{\tau_\varepsilon} - \sum_{t_i^{(f)}}(\varepsilon_i - \delta(t - t^{(f)_i}))$
Yusoffa and Grüning (2012)	$\Delta w_{j,i} = [\alpha + R(t)]\lambda_{j,i}(t)$ $\Delta\lambda_{j,i} = STDP(\Delta(t^{(f)}_{pre/post}))$
Ozturk and Halliday (2016)	$\Delta w_{j,i} = \eta[\bar{R} - R(t)]\lambda_{j,i}$ $\Delta\lambda_{j,i} = \tau_\lambda(STDP(\Delta(t^{(f)}_{pre/post})) - \lambda_{j,i})$

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 **I** 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 $STDP(t_{post} - t_{pre})$, 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).

2.3.3. Approximating the TD error

Q-learning, a family of TD algorithms focused on the optimization of value estimates for pairs of states and actions (Q-values $V_t(S_t, A_t)$ rather than the standard state values $V_t(S_t)$ related to Eqs. (7), (8), and (10)), is a staple of modern RL that addresses both the control (action selection) and evaluation (policy refinement) problems (Sutton & Barto, 2017). For problem domains where function approximation is necessary – typically those tasks involving a continuous rather than discrete set of states and actions – an Actor–Critic approach assigns the problems of control and evaluation to a pair of complimentary neural networks, dubbed "Actor" and "Critic" respectively, that work in an interleaved fashion to optimize the framework's approximation of the true Q-values for the task domain.

Frémaux, Sprekeler, and Gerstner (2013) extended the general framework of R-STDP introduced in Section 2.3.2 to follow this Actor–Critic network design with two networks of SRM_0 spiking neurons. Their formulation of the learning rule for both Actor and Critic neurons replaces the eligibility trace for temporal credit assignment with a smoothing kernel κ whose shape maintains an implicit and decaying record of causal (pre-before-post) paired spiking activity.

$$\Delta w_{j,i} = \alpha D(t)\left(\left[\mathbf{Y}_i(\mathbf{X}^{\hat{t}_i}_j \circ \varrho)\right] \circ \frac{\kappa}{\tau_R}\right)(t) \tag{20}$$

$$D(t) = \frac{R(t)}{N}\left[\sum_{i=1}^{N} \mathbf{Y}_i \circ \left(\kappa' - \frac{\kappa}{\tau_R}\right)(t)\right] - \frac{u_{rest}}{\tau_R} + R(t) \tag{21}$$

$$\kappa = \frac{e^{\frac{-t}{\tau_K}} - e^{\frac{-t}{\vartheta_k}}}{\tau_K - \vartheta_K} \tag{22}$$

Eqs. (20)–(22) describe the dynamics of weight updates for synapses connecting pre-synaptic units j to post-synaptic neurons i in terms of composite decaying learning rate α, TD error estimate $D(t)$, excitatory post-synaptic potentiation dynamics modeled by ϱ (formulation omitted for relevance), kernel κ (and its similarly omitted derivative κ'), kernel decay and rise temporal constants τ_K and ϑ_K, and pre- and post-synaptic spike trains $X^{\hat{t}_i}_j$ and Y_i of the form $\sum \delta(t - t^{(f)}_k)$ as in previous formulae. Note that the pre-synaptic spike train vector $X^{\hat{t}_i}_j$ is restricted only to spikes by pre-synaptic unit j that occurred prior to the most recent spike by post-synaptic neuron i, as signified by the superscript $\hat{t}_i$ which denotes the time of the last spike by i.

Neural units in both the Actor and Critic networks in the continuous control navigation tasks tested in Frémaux et al. (2013) received identical inputs from "place cells" whose spiking activity signals the location of the agent relative to the centers of discrete blocks in the state space. Both Actor and Critic spiking neurons employed the same weight update mechanism outlined above. Units within the Actor population received lateral connections between neurons indicating a preference to navigate in similar directions wherein each unit additionally potentiated those in its neighborhood of action preference and inhibited those whose spikes signal an incompatible choice. Combined with population vector coding, this scheme allowed a discrete number of neural units to encode continuous action choices via N-winner-takes-all action selection.

Frémaux et al. (2013) reported some encouraging experimental results, particularly in the learned behavior of Critic population neurons resembling that of biological dopaminergic "ramp" cells which have been observed to increasingly fire action potentials upon approach to an expected reward. This similarity to spiking activity in biological ramp cells is shown in Fig. 2. Further, their derivations illustrating the processes by which the TD error approximation is backpropagated for learning (despite being essentially undetectable within the observed spiking behavior of individual artificial neural units) hints toward the biological plausibility of some form of distributed backpropagation of value error under R-STDP methods that has, as yet, failed to be directly detected by neuroscientific research but is widely theorized to occur in biological reward-based learning under the reward-prediction error hypothesis.

3. Exploration: Beyond credit assignment

In the context of RL, exploration is the process by which an agent samples both the environment in which it operates and the available action space through which it may alter its relationship with that environment. This sampling process, when combined with a mechanism for estimating the long-term value of states (or state–action pairs in Q-learning), is used to enable exploitative action selection strategies for maximizing the quantity of reward received by the agent over time. This forms the conceptual basis for trial-and-error learning in computational RL theory.

While the value estimation performed by TD methods is highly efficient for dense and static reward landscapes under even very simple exploration strategies such as random search, the exploration process becomes a significant bottleneck in more complex environments. These are often not densely filled with reward-generating states and in some cases the reward generated in response to actions taken in a given state may change or disappear altogether. Learning exploitative action strategies in environments such as these then requires substantially increased sampling of the action and environment spaces, which has obvious impacts to both the temporal and computational efficiency of learning to perform tasks in these environments.

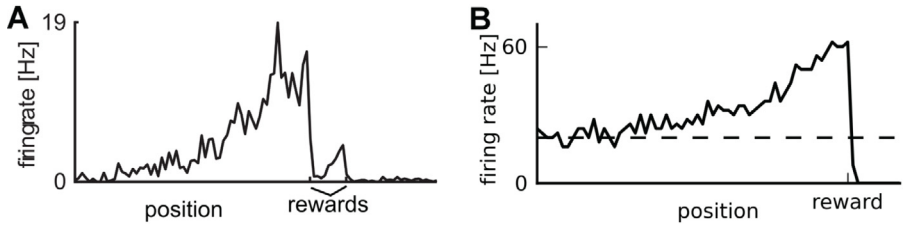

Fig. 2. Illustration of firing rate dynamics for A) rat ventral striatum "ramp cells" during a maze navigation task involving food rewards and B) single Critic population neuron during a linear track task.
Source: Adapted from Frémaux et al. (2013).

The reviewed material in prior sections has focused on the construction of modern neo-Hebbian RL methods which extend biologically plausible forms of Hebbian unsupervised learning to implement neural credit assignment through some mechanism of synaptic selectivity, ranging from tabular traces and short-term weights to more compact kernel methods. While a number of these works briefly consider the issue of stimulating exploration in neo-Hebbian RL agents, their methods for doing so are loosely equivalent to the semi-random search often used as a baseline in RL algorithms. The most common method used in general neo-Hebbian RL is to insert an additive noise function into the calculation of membrane potentials or weight updates to avoid consistently favoring exploitative action choices during learning (Huang, Wu, Yin, & Qiao, 2017).

This section considers works which have extended three-factor neo-Hebbian RL with the intention of developing exploratory behaviors in learning agents that are more capable of adapting to non-trivial reward landscapes, which may be sparse and dynamic.

3.1. Intrinsically motivated reinforcement learning

The concept of intrinsic motivation of behavior and its distinction from extrinsic motivators stems from the study of animal and human behavioral learning in (extrinsically) value-neutral settings (Baldassarre, et al., 2014). The prevailing hypothesis prior to the conceptualization of intrinsic motivators was that biological behavioral learning occurred following a drive to reduce some physical need relative to the survival and reproduction of an organism (Hull, 1943).

If we take physical hunger as an example of one such driver of behavioral learning, then we may view directed foraging as a behavior reinforced by consequent success in reducing the animal's need for nourishment under this theory of drive reduction. While this theory sufficed for experimental evidence around animal conditioning and goal-directed learning, it failed to account for much of the observed behaviors in situations lacking an obtainable goal or reward within the external environment of the organism.

The use of the terms "intrinsic reward" and "intrinsically-motivating" behavior was popularized in the literature on behavioral and developmental psychology by Harlow (Harlow, Harlow, & Meyer, 1950), among others, in attempt to explain behaviors through which organisms expend time and physical effort (finite resources from an evolutionary perspective) with no apparent goal or external benefit. Taking example from the study cited above, rhesus monkeys were shown to learn to efficiently solve non-trivial puzzles in a controlled environment without external incentives such as food. Further, the introduction of extrinsic rewards (treats) during the learning process was found to disrupt rather than enhance performance, leading Harlow to propose that a "manipulation drive" may account for the reinforcement of their puzzle-solving behaviors. Similar drives were proposed across the literature to account for a number of human and non-human behaviors associated with concepts such as curiosity, play, and investigation or exploration. These abstract drivers of behavior are called intrinsic as they lack any direct and discernible connection to an external goal or reward.

While the study of intrinsic motivation has a well-established history in the domain of psychology, its introduction to RL theory is more recent. In Schmidhuber (1991), the issue of efficient exploration by artificial agents led the researchers to propose an RL agent architecture with an adaptive world model. The intent of their design was to produce a system that could dynamically identify poorly modeled portions of the environment space which could then be targeted by a greedy exploration policy to expedite the learning process — essentially a form of meta-learning aimed at autonomously identifying where exploration would be most informative. This was accomplished by jointly training their world model with a "confidence module". This module was trained through supervised gradient methods to approximate the magnitude of weight changes induced in the world model during its training. These approximated improvement values were then used as reinforcement for maximization by the control module of their system. Through learning to estimate the performance improvements made in the world model via sampling of state space transitions, this confidence module was shown to significantly improve the efficiency of training the world model on a given environment.

This approach to incorporating intrinsic motivators in computational RL was formalized in Barto, Singh, Chentanez, et al. (2004) and Chentanez, Barto, and Singh (2004), which focused on hierarchical skill learning. Although their framework was based on option theory and utilized a predetermined measure of salience as additional reinforcement, their work served to illustrate the utility of intrinsic RL for acquiring complex action sequences to attain sparse extrinsic rewards. Schembri, Mirolli, and Baldassarre (2007) expanded upon the methods of Chentanez et al. (2004) both to generalize this approach to handle continuous state and action spaces as well as to autonomously generate additional intrinsic reinforcement signals. These signals were produced via a neural network trained by an evolutionary algorithm and supplemented a simple prediction error calculation to form a surprise-based intrinsic motivation value.

Singh, Lewis, Barto, and Sorg (2010) rigorously assessed the evolutionary significance of intrinsic motivation for RL in the context of optimal reward functions. Their experiments proved that the optimal reward function for a given agent, which may utilize intrinsic rewards to reinforce intermediary behaviors in addition to the primary extrinsic rewards, will outperform or at least equal the performance of the same agent that uses only a fitness-based reward function. Fitness, in this context, relates to the ability of a learning agent to achieve goals, such as successful hunting by wild animals or cumulative point acquisition in the context of a game.

The logic for their theoretical assessment of motivation in the context of evolutionary fitness is straightforward and bears

weight on the use of intrinsic motivation in RL. An entirely fitness based reward function, corresponding to the case of extrinsic-only RL, only rewards behaviors illustrated via experience to confer fitness for the task at hand. Reward functions that additionally account for other factors, such as intrinsic motivators, reinforce intermediate behaviors that may enhance fitness at some later point for the agent — in addition to the same reinforcements provided by the fitness component of the reward function. This does not imply that every possible type and strength of intrinsic reinforcement will produce equal or improved performance compared to fitness-only reinforcement but rather that any given fitness function can be improved upon by balancing it against one or more other factors that incidentally enhance cumulative fitness. How to construct such an improved reward function in the general case remains an open problem in computational RL, and it is possible that successfully managing the exploration–exploitation balance requires a sufficiently good approximation of optimal reward functions.

3.2. Acetylcholine and R-STDP

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.

3.3. Weight saturation and network reconfiguration

A number of rate-based approaches to the exploration-exploitation balance have been proposed in recent neo-Hebbian RL frameworks. These efforts have largely avoided attempts at explicitly modeling additional neuromodulatory factors. One exception is that of Lew, Rey, and Zanutto (2013) which proposed a dual modulation method modeling norepinephrine rather than acetylcholine to alter the excitability of dopaminergic neurons such that exploitative strategies are only favored during periods

of heightened performance (in terms of reward received). While the model produced for that work incorporated network modules and connectivity patterns with robust biological plausibility, their approach to the exploration–exploitation balance can be reduced to a form of semi-random search. This is due to the fact that during periods of lower performance norepinephrine was modeled as having an inhibitory effect on dopaminergic neurons as well as neurons in the input–output response pathway. This resulted in a heightened noise to signal ratio for response pathway neurons such that their output would fail to meet a pre-programmed threshold for activation under a winner-take-all mechanism. When the model failed to produce a response, pre-programmed responses were induced with a predetermined probability. The approach taken in Lew et al. (2013) could be improved by allowing the interactions of dopamine and norepinephrine to alter the threshold value for the winner-take-all output mechanisms during periods of exploration stimulated by norepinephrine under poor performance.

Legenstein, Chase, Schwartz, and Maass (2010) proposed an Exploratory-Hebbian (E-H) learning formulation for learning under dynamic RL tasks. Their approach combined averages of pre- and post-synaptic activations with a low-pass filter to adjust weights such that only rewards above the mean result in reinforcement of coincident activity between connected units. To stimulate exploratory behavior in the model, action selection neurons were provided input from a parameterized source of random noise drawn from a distribution with variance v — the authors term this value as the exploration level parameter. The rate-based E-H weight update formulation is shown below.

$$\Delta w_{j,i} = \eta y_j(t)(y_i(t) - \bar{y}_i(t))(R(t) - \bar{R}(t)) \tag{26}$$

Eq. (26) replaces the direct use of post-synaptic activity in standard Hebbian plasticity with a measure on its deviation from the previous activity level mean $\bar{y}_i$, performing a simple filter on the post-synaptic excitation akin to a varying threshold separating LTP from LTD. The modulation factor corresponding to reward value is similarly filtered against its mean value $\bar{R}$. This was shown by Legenstein et al. (2010) to implicitly perform appropriate credit assignment without the use of eligibility tracing or short-term plasticity components as employed by the methods in Section 2.3.

These approximations of sliding thresholds allow for a number of mechanics for network reconfiguration at the weights. By the term "sliding", we refer to the dynamic nature of this threshold used for induction and reversal of LTP/LTD in contrast with the use of fixed threshold parameters. For example, an above average reward coincident with below average post-synaptic activity triggers LTD, while a below average reward in the same scenario results in LTP. When received rewards increase, neural units which also recently increased their activity should be given the credit for their role in attaining that heightened reward value, therefore this rule enhances the strength of their incoming synaptic weights. Those which reduced their activity prior to an increase in external reward are subject to weakening of the weighting of their input, as the input from neuron j led to reduced activity for unit i when an increase in excitation would have been appropriate. Similar arguments for the case of reduced rewards relative to past averages should be simple to assess here.

While their approach deviates little from the baseline additive noise methods used in standard neo-Hebbian RL, the ability to vary the noise level to stimulate exploratory behavior during learning is a significant factor to consider as we approach more complex methods for biasing exploration in Hebbian RL agents. The value for the variance parameter v need not necessarily be a parameter for human specification, as it was found to function quite well over a broad range of values and only required additional weight normalization for extreme values or when employed alongside an aggressive learning rate η. An interesting variation on the E-H rule could relate the exploration level value v to the deviation of neural activity from recent means, deviation of received rewards from recent means, or both — we explore the potential of such alterations more fully in later sections, where biological plausibility implications may inform their consideration.

Soltoggio and Stanley (2012) introduced Reconfigure-and-Saturate (RaS) Hebbian plasticity. In this work, focus was placed on the role of neuromodulation as a gating mechanism in neo-Hebbian plasticity through adaptive balancing of noise in neural activations and saturation of weight values. In their framework, allowing for the weight values to saturate (up to some maximal value) through typical neo-Hebbian reinforcement was shown to enhance the stability of the network dynamics, giving rise to purely exploitative behavioral strategies. Conversely, a combination of negative reward values and increased noise in signal transmission between connected units served to reset the learned parameters, which trended towards common values and oscillated around them for the duration of the negative signal (implicitly un-learning under negative external value). Trending towards a state of network reconfiguration, in conjunction with noise neural activity, stimulated exploration in a similarly randomized way to previously reviewed works but with the potential for selective re-learning under the exploratory regime, as demonstrated by the experiments in Soltoggio and Stanley (2012).

$$\Delta w_{j,i}(t) = w_{j,i}(t-1) + \big(C \cdot R(t) y_i(t) y_j(t-1)\big) + \xi_{j,i}(t) \tag{27}$$

Eq. (27) shows the weight update formulation corresponding to RaS Hebbian learning. As their framework incorporated learning for both excitatory and inhibitory rate-based neuron types, the variable C takes the value of $+1$ or -1 for each type. This framework additionally modeled causality through explicit propagation delays, with the learning rule relating the activity of the pre-synaptic unit at time $t-1$ with the current post-synaptic activation at time t. ξ is an additive noise to the weight calculation drawn from a uniform distribution for each synapse during updates. Through a series of experiments simulating learning under the proposed plasticity rule, network reconfiguration under LTD induced by negative rewards was found to selectively apply to connections corresponding to behavioral outcomes that required change for successful task progress. One such example discussed was re-learning to navigate due to changes in the environment's reward structure. Previously learned responses to stimuli remained intact where appropriate for attaining reward. This work was the only reviewed neo-Hebbian learning study which forewent the use of eligibility tracing or any comparable mechanism for synaptic credit assignment under delayed reward. As such, future work may consider expanding upon their methods to investigate its performance in more complex tasks and environments.