Desiderata for normative models of synaptic plasticity

Colin Bredenberg[1,2,†] and Cristina Savin[1,3]

[1] Center for Neural Science, New York University, New York, NY 10003, USA
[2] Mila- Quebec AI Institute, 6666 Rue Saint-Urbain, Montréal, QC H2S 3H1
[3] Center for Data Science, New York University, New York, NY 10011, USA
[†]Corresponding author: colin.bredenberg@mila.quebec

Keywords: computational neuroscience, learning, synaptic plasticity

Abstract

Normative models of synaptic plasticity use a combination of mathematics and computational simulations to arrive at predictions of behavioral and network-level adaptive phenomena. In recent years, there has been an explosion of theoretical work on these models, but experimental confirmation is relatively limited. In this review, we organize work on normative plasticity models in terms of a set of desiderata which, when satisfied, are designed to guarantee that a model has a clear link between plasticity and adaptive behavior, consistency with known biological evidence about neural plasticity, and specific testable predictions. We then discuss how new models have begun to improve on these criteria and suggest avenues for further development. As prototypes, we provide detailed analyses of two specific models – REINFORCE and the Wake-Sleep algorithm. We provide a conceptual guide to help develop neural learning theories that are precise, powerful, and experimentally testable.

1 Introduction

Our identities change with time, gradually reshaping our experiences. We remember, we associate, we learn. However, we are only beginning to understand how changes in our minds arise from underlying changes in our brains. Of the many features of neural architecture that are altered over time, from the biophysical properties of

individual neurons to the creating or pruning of synapses between neurons, changes in the strength of existing synapses have long been among the most prominent candidates for the neural substrate of longitudinal perceptual and behavioral change, because many synaptic connections are easily modified, and these modifications can persist for extended periods of time (Bliss and Collingridge, 1993). Further, synaptic modification has been associated with many of the brain's critical adaptive functions, including memory (Martin et al., 2000), experience-based sensory development (Levelt and Hübener, 2012), operant conditioning (Ohl and Scheich, 2005; Fritz et al., 2003), and compensation for stroke (Murphy and Corbett, 2009) or neurodegeneration (Zigmond et al., 1990). However, beyond these associations, a precise link between plasticity and adaptive behaviors of interest is currently lacking.

Here, we distinguish 'normative' modeling approaches from other alternatives, demonstrate why they show promise for establishing this link, and outline a set of desiderata which articulate how recent progress on normative plasticity models strengthens the link between plasticity and system-wide adaptive phenomena. To provide concrete examples of these principles in action, in Appendices C and D we provide worked tutorials on two complementary canonical normative plasticity models—REINFORCE (Williams, 1992) for reinforcement learning, and the Wake-Sleep algorithm for unsupervised learning (Dayan et al., 1995; Hinton et al., 1995)—and illustrate their successes and failures to match our desiderata.

1.1 Phenomenological, mechanistic, and normative plasticity models

We distinguish between three partially overlapping types of model: phenomenological, mechanistic, and normative (Fig. 1a) (Levenstein et al., 2020). The focus of this review is normative plasticity models, but to understand their importance, we first describe their relationship to their counterparts.

In the simplest terms, a phenomenological model's focus is on describing experimental data: the primary goal is to concisely summarize relationships between observed variables. As an example, many early studies of spike-timing-dependent plasticity (STDP) described the relationship between plasticity and the relative timing of pre- and post-synaptic spikes with exponential curves fit to data (Zhang et al., 1998; Dan and Poo, 2004; Sjöström et al., 2010). Such models can reduce the complexity of data, providing interpretability and, to some extent, predictive power. They are incomplete descriptions of the biophysical processes that form the causal link between spike times and plasticity, but extract and summarize important features of the data on which subsequent theories and models can build.

A mechanistic model attempts to explain a set of experimental results in terms of causal interactions between biophysical quantities. For instance, since the initial characterization of STDP, a plethora of studies have emerged characterizing in detail the interactions between backpropagating action potentials (Magee and Johnston, 1997), dendritic morphological properties (Froemke et al., 2005; Letzkus et al., 2006;

Sjöström and Häusser, 2006), local membrane voltage, NMDA ion channel properties, and calcium-sensitive molecules near the synapse. Mechanistic models (Graupner and Brunel, 2010) characterize how these variables all collectively contribute to the strengthening or weakening of the synapse. As a consequence of their depth and breadth, mechanistic models can often provide predictions that are outside of the scope of the original experiment, and provide useful targets for experimental manipulation.

The distinction between phenomenological models and mechanistic models is not always completely crisp, especially in areas where our scientific understanding is progressing rapidly. In nascent mechanistic models, there often exist 'black boxes' that specify interactions between known biophysical quantities, without a precise understanding of whether or how these interactions are implemented (Craver, 2007). Because they lack a direct relation to well-understood biophysics, these 'black boxes' act in essentially the same way as variables do in a phenomenological model. In this way, we can see that there exists a spectrum between phenomenological and mechanistic models, and that oftentimes, mechanistic models grow from phenomenological ones. However, there is more to the spectrum: while phenomenological and mechanistic models articulate how synaptic plasticity works, they do not explain *why* it exists in the brain, i.e. what its importance is for neural circuits, behavior, or perception. To answer this question with any precision requires an appeal to normative modeling.

Normative models aim answer this 'why' question by connecting plasticity to observed network-level or behavioral-level phenomena, including memory formation (Hopfield, 1982) and consolidation (Benna and Fusi, 2016; Clopath et al., 2008; Fusi et al., 2005), reinforcement learning (Frémaux and Gerstner, 2016), and representation learning (Hinton et al., 1995; Oja, 1982; Rao and Ballard, 1999; Savin et al., 2010). This class of plasticity model, in our view, employs a fundamentally different set of methodologies from phenomenological or mechanistic models, in order to provide the missing link between plasticity and function. Guided by the intuition that plasticity processes have developed on an evolutionary timescale to near-optimally perform adaptive functions, normative plasticity theories are typically 'top-down', in that they begin with a set of prescriptions about how synapses 'should' modify in order to optimally perform a given learning-based function. Subsequently, with varying degrees of success, these theories attempt to show that real biology matches or approximates this optimal solution. As an example, an increasing body of literature is establishing a correspondence between classical reinforcement learning algorithms (Williams, 1992) and reward-modulated Hebbian synaptic plasticity models of learning in the brain (Frémaux and Gerstner, 2016). This process is ongoing, and though experimental support for such forms of plasticity are growing (Gerstner et al., 2018), much work remains to be done. Similar efforts are underway to construct approximations to the backpropagation algorithm which can serve as models of neural plasticity (Marschall et al., 2020; Lillicrap et al., 2020; Richards and Lillicrap, 2019; Urbanczik and Senn, 2014). Here, we will review classical normative plasticity approaches and discuss recent efforts to improve upon them.

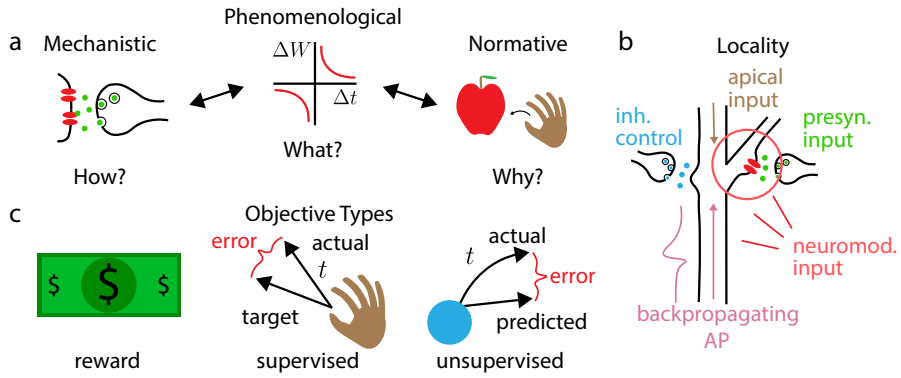

Figure 1: **Defining normative modeling. a.** Spectrum of synaptic plasticity models. Mechanistic models show how detailed biophysical interactions produce observed plasticity, phenomenological models concisely describe what changes in experimental variables (e.g. post-pre relative spike timing Δt) affect plasticity (ΔW), and normative models explain why the observed plasticity implements capabilities that are useful to the organism. **b.** Schematic illustrating the range of local variables that may be available for synaptic plasticity. These include, but are not limited to: backpropagating action potentials from the soma, apical dendritic input, pre- and postsynaptic activity, neuromodulatory signals, and potentially inhibitory input from local microcircuitry. **c.** Classes of objective function used in normative plasticity theories. Reward-based objectives involve only feedback about how well the organism or network performed, whereas supervised objectives provide explicit targets for network output. By contrast, unsupervised objectives do not require any form of explicit feedback to train the network.

2 Desiderata for normative models

One of the biggest challenges for a normative model of synaptic plasticity is its connection to biology: artificial neural networks with simulated synapses (synaptic weight parameters) that adapt to improve performance on any of a variety of functions from sensory processing (LeCun et al., 1989; Krizhevsky et al., 2012), to motor learning (Heess et al., 2017; Hafner et al., 2019), to abstract game learning (Silver et al., 2017; Vinyals et al., 2019) are much more accessible to mathematical and empirical investigation than the neural circuits implementing these functions in the brain. Compared to the simulations and mathematical analysis used to explore machine learning algorithms, neuroscience experiments are time-consuming and expensive. Further, network simulations provide total access to neural activations, stimuli, and synaptic parameters over the whole course of learning, whereas any one neuroscience experiment can only reveal a very small amount about what is going on in a circuit. Therefore, it is a major challenge to identify how to improve normative models with relatively limited access to experimental data confirming or rejecting their predictions.

In what follows, we will articulate a set of desiderata that can serve as both an

organizing tool for understanding the contributions of recent normative plasticity modeling efforts and as intermediate objectives for the development of new models in the absence of explicit experimental rejection or confirmation of older work. We will argue that each principle is desirable for some combination of the following reasons: first, it may help ensure that the plasticity model actually qualifies as normative; second, it may require a model to accommodate known facts about biology; third, it may ensure that models can be compared properly to existing experimental literature and generate genuinely testable experimental predictions. Most of these desiderata are relatively intuitive and simple. However, it has proven incredibly difficult for existing models of any adaptive cognitive phenomenon—from sensory representation learning, to associative memory formation, to reinforcement learning—to satisfy all desiderata in tandem.

2.1 Improving performance

One way to view the normative approach is that it attempts to organize the diversity of synaptic dynamics existing within a neural system into the simplest explanatory framework possible for what functions the system's plasticity subserves. Usually, this framework is mathematical for pragmatic reasons: mathematics provides the precision and power necessary to establish clear relationships between plasticity and function. In particular, viewing neural plasticity as an approximate optimization process has been very fruitful (Lillicrap et al., 2020; Richards et al., 2019), wherein synaptic modifications progressively reduce a scalar loss function. This process can be divided into two steps: articulating an appropriate objective, and subsequently demonstrating that a synaptic plasticity mechanism improves performance on that objective.

It can be extremely difficult to reduce the full range of functions a given circuit must perform to a scalar objective function, but as we will show subsequently, the conceptual benefits can be immense. On one side, picking too simple an objective function runs the risk of ignoring many functions a system is required to perform. For instance, early normative theories of learning in sensory systems show how synaptic plasticity could minimize the objective function underlying principal component analysis (PCA) (Oja, 1982), but merely representing the principal components of an incoming sensory stream is an inadequate characterization of sensory processing for several reasons. PCA can capture only second-order properties (mean and covariance) of naturalistic stimuli and does not perform the highly nonlinear processing required for cortical neurons to exhibit gain control capabilities (Simoncelli and Heeger, 1998) and texture (Ziemba et al., 2016) and object class (Rust and DiCarlo, 2010) selective responses. A given synaptic plasticity mechanism may only be able to minimize a restricted subset of objectives, and for a normative theory, the set of possible objectives that can be minimized must encompass a wide range of functions that the brain is known to subserve. Beyond principal component analysis, many modern models of unsupervised representation learning use objectives for training hierarchical generative models (e.g. the evidence lower bound (ELBO) which underlies the Wake-Sleep algorithm (Dayan et al., 1995) and predictive coding (Rao and

Ballard, 1999), and allows for multilayer, nonlinear representation learning). On the other side, selecting too flexible an objective function can run the risk of 'overfitting' experimental data, a problem that is particularly salient for Bayes-optimal accounts of neuroscientific and psychological phenomena (Bowers and Davis, 2012). As an extreme example, if we were to postulate that the 'objective' of a neural system is to behave exactly as it is observed to behave experimentally, i.e. everything in a neural system happens precisely as was 'intended', then the normative project becomes vacuous: the model provides neither conceptual simplification nor predictive power beyond what was observed experimentally, and has consequently failed to provide a useful explanation of the data. Therefore, the quality of an objective function is determined by both how many phenomena it is able to explain and how simple it is.

Normative theories of synaptic plasticity developed to date usually involve some combination of supervised, unsupervised, or reinforcement learning objectives (Fig. 1c). The choice of objective function for a neural system is laden with philosophical assumptions about the system's functional utility, and can exert a huge influence on the resultant form and scope of applicability of the synaptic plasticity model. For instance, supervised learning usually involves the existence of either an internal or external teacher. If the teacher is external, such a learning mechanism could only be leveraged under the very specific and comparatively rare conditions in which the organism is being overtly taught, as is the case, for instance, in some models of zebra finch song learning (Fiete et al., 2007). If the teacher is internal, a plausible normative theory is limited in the types of knowledge the 'self-supervisor' may reasonably construct and provide (for instance, motor error signals (Gao et al., 2012; Bouvier et al., 2018) or saccade information indicating that a visual scene has changed (Illing et al., 2021)). Generative modeling is a form of unsupervised learning that postulates that a sensory system is actively building a probabilistic model of its sensory inputs, which can be used to simulate possible future outcomes and perform Bayesian reasoning (Fiser et al., 2010). This vision of sensory coding is popular both for its ability to accommodate normative plasticity theories (Rao and Ballard, 1999; Dayan et al., 1995; Kappel et al., 2014; Bredenberg et al., 2021) and for its philosophical vision of sensory processing as a form of advanced model building, beyond simple sensory transformations. However, model construction is only indirectly useful for many tasks involving rewards and planning, and so such plasticity would have to occur concomitantly with reward-based (Frémaux and Gerstner, 2016) or motor (Gao et al., 2012; Feulner and Clopath, 2021) learning. Furthermore, alternative perspectives on sensory processing exist, including those based on maximizing the information about a sensory stimulus contained in a neural population (Attneave, 1954; Atick and Redlich, 1990) subject to metabolic efficiency constraints (Tishby et al., 2000; Simoncelli and Olshausen, 2001), and those based on 'contrastive methods' (Oord et al., 2018; Illing et al., 2021), where a self-supervising internal teacher encourages the neural representation of some stimuli to grow closer together, while encouraging others to grow more discriminable.

Evaluating which objective function (or functions) best explains the properties of a neural system is very hard: while some forms of objective function may have discriminable effects on plasticity (e.g. supervised vs. unsupervised learning (Nayebi et al.,

2020)), others are even provably impossible to distinguish. As a simple example, suppose that we have an N^r dimensional single-layer neural network receiving N^s dimensional stimuli through an $N^r \times N^s$ dimensional weight matrix $\mathbf{W}$. We have the response given by:

$$\mathbf{r} = f(\mathbf{W}\mathbf{s}), \tag{1}$$

where $f(\cdot)$ is a tanh nonlinearity. Now suppose that some setting of synaptic weights $\mathbf{W}^*$ minimizes an objective function $\mathcal{L}$, i.e. $\mathcal{L}(\mathbf{W}^*) \leq \mathcal{L}(\mathbf{W})\ \forall \mathbf{W}$. We might be tempted to argue that because $\mathbf{W}^*$ minimizes $\mathcal{L}$, $\mathcal{L}$ must be the objective that the system is minimizing. However, there are an infinite variety of alternative objectives that share this same minimum (Appendix A). This motivates the idea that for a given dataset, it is very plausible that one objective ($\tilde{\mathcal{L}}$) can *masquerade* as another ($\mathcal{L}$). In some cases, complex objective functions can masquerade as simple objectives, which may only be epiphenomenal. For instance, it has been hypothesized that synaptic modifications may preserve the balance between inhibitory and excitatory inputs to a cell (Vogels et al., 2011); recent theories have proposed that this E/I balance may only be a consequence of a more advanced theory of sensory predictive coding (Brendel et al., 2020). In other cases, philosophically distinct frameworks, such as generative modeling, information maximization, or denoising may simply produce similar synaptic plasticity modifications because the frameworks often overlap heavily (Vincent et al., 2010), and may not be distinguishable on simple datasets without targeted experimental attempts to disambiguate between the two perspectives.

Furthermore, not every function performed by biological systems has been adequately incorporated into a simple optimization framework. For example, though the Hebbian plasticity rule used in Hopfield networks endows model circuits with associative memory, the utility of learning is characterized by the dynamical attractor structure it embeds in the neural circuit, rather than by its direct minimization of an objective function (Hopfield, 1982). In addition, the notion that some parts of the brain may have synaptic plasticity mechanisms for representation learning while other parts have plasticity for reinforcement learning suggests that the brain may be better viewed as a collection of interacting systems with only partially overlapping goals. This multiagent (Zhang et al., 2021) formulation of learning has intuitive appeal, because it can decompose broad objectives like survival into a series of intermediate objectives carried out by individual systems. Such a formulation could help explain how locality emerges, i.e. why synapses do not need information about distant neural circuits in order to improve performance. However, with this additional appeal comes additional conceptual and mathematical complexity, because improving performance on one objective could very easily harm the performance of other systems. Therefore, insofar as a collection of neural circuits and plasticity mechanisms *can* be viewed as acting in concert to improve a unified objective, simple optimization is the preferable perspective.

Having addressed many difficulties associated with choosing a good objective function, we now move to difficulties involved in demonstrating that a particular synaptic plasticity rule decreases a chosen objective[1]. How could such a property be proven?

[1]Some objectives (like reward functions) are best thought of as being maximized rather

For a particular plasticity rule to reduce an objective, we need to show that the following principle holds:

$$\begin{aligned} &\mathcal{L}(\mathbf{W} + \Delta\mathbf{W}) < \mathcal{L}(\mathbf{W}) \\ \Rightarrow &\mathcal{L}(\mathbf{W} + \Delta\mathbf{W}) - \mathcal{L}(\mathbf{W}) < 0, \end{aligned} \tag{2}$$

for some update $\Delta\mathbf{W}$ determined by the plasticity rule. If we accept the additional supposition that $\Delta\mathbf{W}$ is very small, we can employ the first order Taylor approximation (treating $\mathbf{W}$ as a flattened vector of length $N^r \times N^s$): $\mathcal{L}(\mathbf{W} + \Delta\mathbf{W}) \approx \mathcal{L}(\mathbf{W}) + \frac{d\mathcal{L}}{d\mathbf{W}}(\mathbf{W})^T \Delta\mathbf{W}$. Substituting this approximation into our reduction criterion, we have after cancellation:

$$\frac{d\mathcal{L}}{d\mathbf{W}}(\mathbf{W})^T \Delta\mathbf{W} < 0. \tag{3}$$

This shows that for small weight updates (slow learning rates), the inner product between a synaptic learning rule $\Delta\mathbf{W}$ and the gradient of the selected loss function $\mathcal{L}(\mathbf{W})$ with respect to the weight change must be negative. The simplest way to ensure that this is true is for $\Delta\mathbf{W}$ to equal a small scalar λ times the negative gradient of the loss ($-\lambda \frac{d\mathcal{L}}{d\mathbf{W}}(\mathbf{W})^T \frac{d\mathcal{L}}{d\mathbf{W}}(\mathbf{W}) = -\lambda \| \frac{d\mathcal{L}}{d\mathbf{W}}(\mathbf{W}) \|_2^2 < 0$). If this were true, plasticity would be guaranteed to improve performance on the objective $\mathcal{L}$. Unfortunately, for even the simplest neural networks and objective functions, naive methods of calculating this gradient will prove to be nonlocal (see Appendix B for a simple example). Thus, the critical challenge for normative theories of synaptic plasticity is finding ways that neural networks can find synaptic modifications $\Delta\mathbf{W}$ that demonstrably have a negative inner product with the gradient of a desired objective $\mathcal{L}$, while still allowing the neural network to satisfy biologically realistic locality constraints. However, it is important to note that if an update $\Delta\mathbf{W}$ reduces any one objective function, then it also reduces an infinite number of similar alternative objective functions (Appendix A); therefore it is perhaps best to think of normative plasticity models in terms of the family of objective functions that they minimize—committing to any one particular objective within that family reflects the predilections of the theorist, not the system.

Different normative studies demonstrate that Eq. 3 holds by different methods. Some studies show empirically across many simulations that this inner product is negative (Lillicrap et al., 2016; Marschall et al., 2020). However, these demonstrations alone do not answer the following questions: how would we know that the network would still perform well if a different task were chosen, or if the network's architecture were different, or if various elements of the simulated plasticity mechanism were changed? A simulation has relatively limited power to extrapolate beyond its immediate results, especially when the neuron models used in large-scale network simulations are often very reductive (Gerstner and Kistler, 2002) and when small changes in simulated network parameters can effect large qualitative differences in

than minimized. Without loss of generality, in such cases we can minimize the negative reward function.

network behavior (Xiao et al., 2021). Further, a battery of *in silico* simulations under a variety of different parameter settings and circumstances rapidly begins to suffer the curse of dimensionality, becoming almost as extensive as the collection of *in vivo* or *in vitro* experiments that it is attempting to explain. As such, simulation-based justifications suffer from a lack of conciseness and an inability to easily address counterfactuals.

For this reason, much focus in the field has been devoted to constructing mathematical arguments as to why Eq. 3 should hold for a given local synaptic plasticity rule. Some plasticity rules amount to stochastic approximations to the true gradient (Williams and Zipser, 1989; Scellier and Bengio, 2017) and some are systematically biased but maintain a negative inner product under reasonable assumptions (Bredenberg et al., 2021; Dayan et al., 1995; Amari and Nakahara, 1999; Meulemans et al., 2020). Mathematical analysis allows one to know quite clearly when a particular plasticity rule will decrease a loss function, and identifies how plasticity mechanisms should change with changes in the network architecture or environment. However, analysis is often only possible under restrictive circumstances, and it is often necessary to supplement mathematical results with empirical simulations in order to demonstrate that the results extend to more general, more realistic circumstances.

2.2 Locality

Biological synapses can only change strengths using chemical and electrical signals available at the synapse itself. 'Locality' refers to the idea that a postulated synaptic plasticity mechanism should only refer to variables that could be conceivably available at a given synapse (Fig. 1b). Though locality may seem like an obvious requirement for any theory of biological function, for synaptic plasticity it presents a great mystery: how does a system as a whole, whose success or failure is determined by the joint action of many neurons distributed across the entire brain, communicate information to individual synapses about how to improve? The success of most machine learning algorithms relies on nonlocal, even global, propagation of learning signals, including backpropagation (Werbos, 1974; Rumelhart et al., 1985) (See Appendix B), backpropagation through time (Werbos, 1990), and real-time recurrent learning (Williams and Zipser, 1989).

Despite its importance as a guiding principle for normative theories of synaptic plasticity, locality is a slippery concept, primarily because of our insufficient understanding of the precise battery of biochemical signals available to a synapse, and how those signals could be used to approximate quantities required by theories. As a simple example, many normative theories require information about the pre- and postsynaptic firing rates of a neuron, similar to Hebb's Postulate (Hebb, 1949). However, neurons predominately communicate to one another through discrete action potentials, and additional cellular machinery would be required to form an estimate of pre- and postsynaptical firing rates based on backpropagating action potentials from the soma and on postsynaptic potentials. Whether a plasticity rule derived from normative principles involves rate or spike-based information is often a

function of the neuron model used in the theory, and it is often difficult to formulate predictions about how a realistic, non-idealized neuron should exactly modify its synapses based on over-simplified models. Therefore, normative theories typically declare success when some standard of plausibility is reached, where derived plasticity rules roughly match the experimental literature (Payeur et al., 2021) or only require reasonably simple functions of postsynaptic and pre-synaptic activity that a synapse could hypothetically approximate (Oja, 1982; Scellier and Bengio, 2017; Williams, 1992).

In normative models of synaptic plasticity, the requirement of locality is in perpetual tension with the general requirement for some form of 'credit assignment' (Lillicrap et al., 2020; Richards et al., 2019), i.e. a mechanism capable of signaling to a neuron that it is 'responsible' for a network-wide error, and should modify its synapses to reduce errors. Depending on a network's objective, a system's credit assignment mechanism *could* take a wide variety of forms, some small number of which may only require information about the pre- and post-synaptic activity of a cell (Oja, 1982; Pehlevan et al., 2015, 2017; Obeid et al., 2019; Brendel et al., 2020), but many of which appear to require the existence of some form of error (Scellier and Bengio, 2017; Lillicrap et al., 2016; Akrout et al., 2019) or reward-based (Williams, 1992; Fiete et al., 2007; Legenstein et al., 2010) signal.

The extent to which a credit assignment signal postulated by a normative theory meets the standards of 'locality' depends heavily on the nature of the signal. For instance, there is growing support for the idea that neuromodulatory systems, distributing dopamine (Otani et al., 2003; Calabresi et al., 2007; Reynolds and Wickens, 2002), norepinephrine (Martins and Froemke, 2015), oxytocin (Marlin et al., 2015), and acetylcholine (Froemke et al., 2013; Guo et al., 2019; Hangya et al., 2015; Rasmusson, 2000; Shinoe et al., 2005) signals can propagate information about reward (Guo et al., 2019), expectation of reward (Schultz et al., 1997), and salience (Hangya et al., 2015) diffusely throughout the brain to induce or modify synaptic plasticity in their targeted circuits. Therefore, it may be reasonable for normative theories to postulate that synapses have access to global reward or reward-like signals, without violating the requirement that plasticity be affected only by locally-available information (Frémaux and Gerstner, 2016).

Locality as a desideratum serves as a heuristic stand-in for the requirement that a normative model must be eventually held to the standard of experimental evidence. This is not to say that normative models cannot postulate neural mechanisms that have not yet been observed experimentally. However, for such an exercise to be constructive, the theory should clearly articulate how it deviates from the current state of the experimental field, and how these deviations can be tested (Section 2.7; see Appendices C and D for concrete examples of this process). Furthermore, the process of mathematical abstraction necessitates approximation (Cartwright and McMullin, 1984): constraining a normative theory to adhere to 'locality' without necessarily requiring a perfect correspondence to experimental data allows normative theories to strive to capture the essence of synaptic learning processes without becoming mired in technical details.

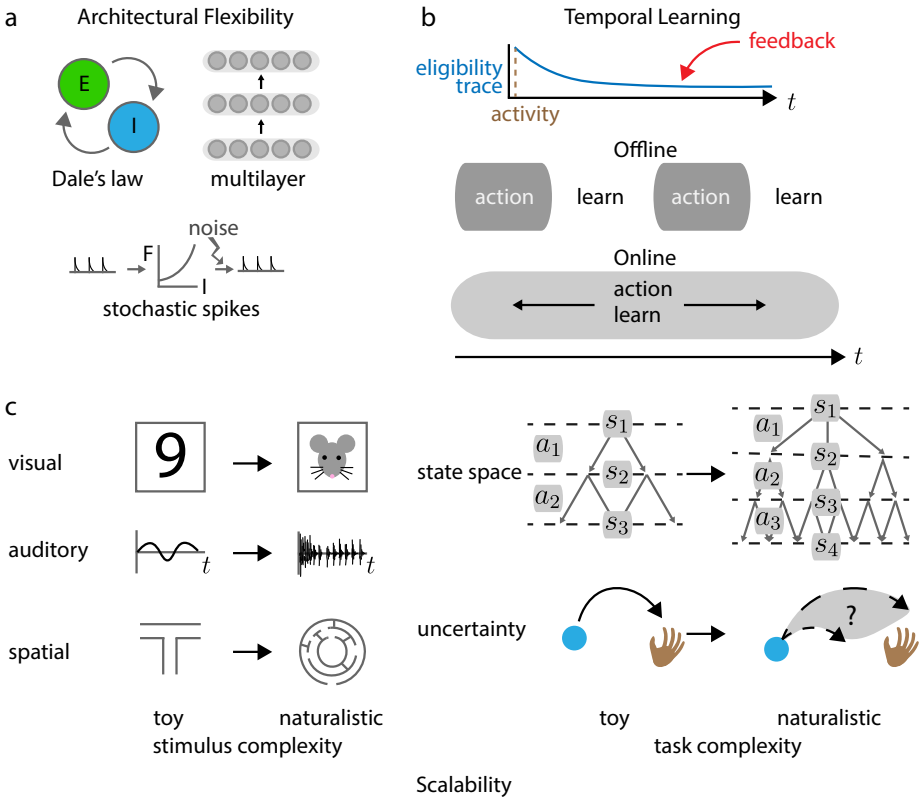

Figure 2: **Architecture and scalability considerations for normative plasticity models. a.** Features of realistic biological networks that normative plasticity theories should be able to account for: separation of excitatory and inhibitory neuron populations; stochastic and spiking input-output functions for individual neurons; and multilayer, recurrent connectivity. **b.** For actions in the past to be associated with delayed supervisory or reinforcement signals, plasticity algorithms require a mechanism of temporal association. One candidate is the 'eligibility trace,' which stores information about coactivity throughout time locally to a synapse, and subsequently modifies synaptic connections when paired with feedback information. Learning can occur offline, where some or all synaptic modification occurs in the absence of action or perception by the organism. Alternatively, it can occur online, where the organism acts and learns simultaneously. **c.** Stimuli (left) and task structure (right) can become complex in many ways. Different sensory features (e.g. visual, auditory, or spatial information) can all be made more naturalistic by training networks on stimuli organisms are exposed to and learn from in natural environments. Further, tasks can be made more naturalistic by increasing the number of action options (a) and sequential state (s) transitions required for a network to achieve its goals and by adding uncertainty into the task.

2.3 Architectural plausibility

The learning algorithm implemented by a plasticity model often requires specific architectural motifs to exist in a neural circuit in order to deliver reward, error, or

prediction signals. These might include diffuse neuromodulatory projections (Fig. S1b) or neuron-specific top-down synapses onto apical dendrites (Fig. S2c). Such architectural features (or alternative, isomorphic motifs) are *required* for the learning algorithm in question, and are known to exist in a wide range of cortical areas. However, normative plasticity models should not depend on circuit features that have been demonstrated not to exist in the modeled system, because spurious architectural features can be used to 'cheat' at achieving locality by postulating unrealistic credit assignment mechanisms (see Appendix B). Further, models lacking important features of neural circuits can be difficult to relate to experimental data. In what follows, we will highlight several particularly important architectural motifs that have been the focus of recent work.

Contrary to the highly reduced deterministic rate-based models typically used in machine learning, neurons communicate through roughly discrete action potentials. Further, they exhibit numerous forms of variability due in part to synaptic failures and constant receipt of task-irrelevant signals (Fig. 2a) (Faisal et al., 2008). Normative theories which employ rate-based activations (Bredenberg et al., 2020; Scellier and Bengio, 2017) or which assume that the input-output function of neurons is approximately linear (Oja, 1982), may not extend to the more realistic discrete, stochastic, and highly nonlinear setting. Further, by ignoring spike timing, such theories inherently produce plasticity rules that ignore the precise relationship between pre- and post-synaptic spike times, and will consequently be unable to capture STDP results. This both limits the expressive power of such models, and prevents their experimental validation. Fortunately, several methods which were originally formulated using rate-based models have subsequently been extended to spiking network models to great effect. Reward-based Hebbian plasticity based on the REINFORCE algorithm (Appendix C) (Williams, 1992) has been generalized to stochastic spiking networks (Frémaux et al., 2013), while backpropagation approximations (Murray, 2019) and predictive coding methods (Rao and Ballard, 1999) have subsequently extended to deterministic spiking networks (Bellec et al., 2020; Brendel et al., 2020). Therefore, a lack of a generalization to spiking networks is not necessarily a death knell for a normative theory, but many existing theories lack either an explicit generalization to spiking or a clear relationship to STDP, and the mathematical formalism that defines these methods may require significant modification to accommodate the change.

Real biological networks have a diversity of cell types with different neurotransmitters and connectivity motifs. At the bare minimum, a normative model must be able to accommodate Dale's Law (Fig. 2a), which stipulates that the neurotransmitters released by a neuron are either excitatory or inhibitory, but not both (for the most part (O'Donohue et al., 1985)). Though this might seem like a simple principle, enforcing Dale's principle can seriously damage the performance of artificial neural networks without careful architectural considerations (Cornford et al., 2021). Furthermore, the mathematical results of *many* canonical models of synaptic modification rely on symmetric connectivity between neurons, including Hopfield networks (Hopfield, 1982), Boltzmann machines (Ackley et al., 1985), contrastive Hebbian learning (Xie and Seung, 2003), and predictive coding (Rao and Ballard,

1999); this symmetry is partially related to the symmetric connectivity required by the backpropagation algorithm (Appendix B). Symmetric connectivity means that the connection from neuron A to neuron B must be the same as the reciprocal connection from neuron B to neuron A. It inherently violates Dale's Law, because it means that entirely excitatory and entirely inhibitory neurons can never be connected to one another: the positive sign for one synapse and the negative sign for the reciprocal connection violates symmetry. Some models, such as Hopfield networks (Sompolinsky and Kanter, 1986) and equilibrium propagation (Ernoult et al., 2020) have been extended to demonstrate that moderate deviations from symmetry can exist and still preserve function. Further, a recent mathematical reformulation of predictive coding has demonstrated that inter-layer symmetric connectivity is not necessary (Golkar et al., 2022). Therefore, recent results indicate that many canonical models believed to depend on symmetric connectivity can be improved upon.

Many early plasticity models, including Oja's rule (Oja, 1982) and perceptron learning (Rosenblatt, 1958), as well as more modern model recurrent network models focused on learning temporal tasks (Murray, 2019) are designed to greedily optimize layer-wise objectives, and their mathematical justifications do not generalize to multi-layer architectures. Though greedy layer-wise optimization may be sufficient for some forms of unsupervised learning (Illing et al., 2021), a method that cannot account for how credit assignment signals are passed between cortical areas will not in general be able to support many complex supervised or reinforcement learning tasks humans are known to learn (Lillicrap et al., 2020). Generalizing layer-local learning to multi-layer objective functions has been the focus of much recent work: many multi-layer models can be seen as generalizations of perceptron learning (Bengio, 2014; Hinton et al., 1995; Rao and Ballard, 1999), with other models such as those derived from similarity matching (Pehlevan et al., 2017) receiving similar treatment (Obeid et al., 2019). We will refer to this form of multi-layer signal propagation as 'spatial' credit assignment, and will refer to relaying information across time as 'temporal' credit assignment (Fig. 2b; Section 2.4). As we will discuss in the next section, models that do not support temporal credit assignment will not be able to account for learning in inherently sequential tasks.

2.4 Temporal credit assignment

Because so many learned biologically-relevant tasks involving temporal decision-making (Gold and Shadlen, 2007) or working memory (Compte et al., 2000; Wong and Wang, 2006; Ganguli et al., 2008) inherently leverage information from the past to inform future behavior, and because neural signatures associated with these tasks exhibit rich recurrent dynamics (Brody et al., 2003; Shadlen and Newsome, 2001; Mante et al., 2013; Sohn et al., 2019), many aspects of learning in the brain require a normative theory of synaptic plasticity that works in recurrent neural architectures and provides an account of temporal credit assignment.

Temporal credit assignment is an important point of failure of modern deep learning

methods, in part due to the inherent instabilities involved in performing gradient descent on recurrent neural architectures (Bengio et al., 1994). That models unconstrained in their correspondence to biology have difficulties handling temporal signals should be some indication of the difficulties posed by temporal credit assignment for normative theories of synaptic plasticity. However, recent improvements in neural architectures, including gated recurrent units (Chung et al., 2014) and long short-term memory units (Hochreiter and Schmidhuber, 1997), as well as sequential reinforcement learning methods (Mnih et al., 2015; Arjona-Medina et al., 2019; Hung et al., 2019; Raposo et al., 2021), have combined to produce several high-profile advances on inherently temporal, naturalistic tasks like game-playing (Silver et al., 2017) and natural language processing (Devlin et al., 2018; Radford et al., 2018). This may indicate that the time is ripe to begin incorporating new developments in deep learning into normative plasticity models.

As it currently stands, the majority of normative synaptic plasticity models focus only on spatial credit assignment, which presents distinct challenges when compared to temporal credit assignment (Marschall et al., 2020). In fact, many theories that provide a potential solution to spatial credit assignment do so by requiring networks to relax to a 'steady-state' on a timescale much faster than inputs (Hopfield, 1982; Scellier and Bengio, 2017; Bredenberg et al., 2020; Xie and Seung, 2003; Ackley et al., 1985), which effectively prevents networks from having the rich, slow internal dynamics required for many temporal motor (Hennequin et al., 2012) and working memory (Wong and Wang, 2006) tasks. Other methods appear to be agnostic to the temporal properties of their inputs, but have not yet been combined with existing plasticity rules that perform approximate temporal credit assignment within local microcircuits (Murray, 2019; Bellec et al., 2020).

While most normative theories focus on spatial credit assignment, some new algorithms do provide potential solutions to temporal credit assignment, through either explicit approximation of real time recurrent learning (Marschall et al., 2020; Bellec et al., 2020; Murray, 2019), by leveraging principles from control theory (Gilra and Gerstner, 2017; Alemi et al., 2018; Meulemans et al., 2022), or by leveraging principles of stochastic circuits that are fundamentally different from traditional explicit gradient-based calculation methods (Bredenberg et al., 2020; Miconi, 2017). Many use what is called an 'eligibility trace' (Gerstner et al., 2018) (Fig. 2b)—a local synaptic record of coactivity—to identify associations between rewards and neural activity that may have occurred much further in the past. We suggest that these models capture something fundamental about learning across time, and that much work remains to combine these with spatial learning rules to construct normative models of full spatiotemporal learning.

2.5 Combining learning and active performance

Similar to the importance of understanding temporal credit assignment in the brain, it is critical to understand how learning in the brain relates to continuous action and perception in an environment (Fig. 2b). The relationship between learning and

active performance in the brain can vary widely depending on the experimental context: learning-related changes can occur concomitantly with action (Bittner et al., 2015; Sheffield et al., 2017; Grienberger and Magee, 2022) ('online' learning), during brief periods of quiescence between trials (Pavlides and Winson, 1989; Bönstrup et al., 2019; Liu et al., 2021), or over periods of extended sleep (Gulati et al., 2017; Eschenko et al., 2008; Girardeau et al., 2009) ('offline' learning). Therefore, whether a normative plasticity model uses offline or online learning should be determined by the experimental context, be it for instance rapid place cell reorganization in new environments, or long timescale memory consolidation.

However, many classical algorithms—especially those that support multi-layer spatial credit assignment (Ackley et al., 1985; Xie and Seung, 2003; Dayan et al., 1995)—are constrained to modeling only offline learning, because they require distinct training phases, during at least one phase of which activity of neurons is driven for *learning*, rather than performative purposes. It has not been clear whether such algorithms are fundamentally offline, or whether the space of phenomena that they can model can be expanded until recently. Some existing two-phase normative algorithms, such as the Wake-Sleep algorithm (Appendix D) (Hinton et al., 1995; Dayan et al., 1995), have be adapted such that the second phase becomes indistinguishable from perception (Bredenberg et al., 2020; Ernoult et al., 2020). Other recent models allow for simultaneous multiplexing of top-down learning signals and bottom-up inputs (Payeur et al., 2021), which enables online learning. These results suggest that future work may fruitfully adapt existing offline algorithms to provide good models of explicitly online learning in the brain.

2.6 Scaling in dimensionality and complexity

A point often underappreciated in computational neuroscience (and possibly overappreciated in machine learning) is that models of learning in the brain need to be able to scale to handle the full complexity of the problems a given model organism has to solve. As obvious as this sounds, it is a point that can be difficult to verify: how can we guarantee that adding more neurons and more complexity will not make a particular collection of plasticity rules more effective? As a case study, consider REINFORCE ((Williams, 1992); Appendix C), an algorithm which, for the most part, satisfies our other desiderata for normative plasticity for the limited selection of tasks in naturalistic environments which are explicitly rewarded. However, though REINFORCE demonstrably performs better than its progenitor weight perturbation (Jabri and Flower, 1992), as the dimensionality of its stimuli, the number of neurons in the network, and the delay time between neural activity and reward increases, the performance of the algorithm decays rapidly, both analytically and in simulations (Werfel et al., 2003). This is primarily caused by high variance of gradient estimates provided by the REINFORCE algorithm, and is only partially ameliorated by existing methods that reduce its variance (Bredenberg et al., 2021; Ranganath et al., 2014; Mnih and Gregor, 2014; Miconi, 2017). Thus, adding additional complexity to the network architecture actually *impairs* learning.

We do not mean to imply that all normative plasticity algorithms should be demonstrated to meet human-level performance, or even that they should match state-of-the-art machine learning methods. Machine learning methods profit in many ways from their biological implausibility: they use stochastic backpropagation, which is demonstrably biologically implausible (Appendix B) but which benefits from very low variance gradient estimates (Werfel et al., 2003); they share weights across topographically distant space in convolutional neural networks (Fukushima and Miyake, 1982); they use rate-based units, which generally perform better than spiking units (Neftci et al., 2019); and they are usually deterministic, which obviates the need for redundancy (increased neuron numbers) and increased computational demand. Beyond machine learning methods, the human brain itself has orders of magnitude more neural units and synapses than have ever been simulated on a computer, all of which are capable of processing totally in parallel. Therefore, direct comparison to the human—or any—brain is also not fair. We propose the far softer condition that as the complexity of input stimuli and tasks increase, within the range supported by current computational power, plasticity rules derived from normative theory should continue to perform well both in simulation and, preferably, analytically. Further, the performance of normative plasticity algorithms can fruitfully be compared to existing machine learning methods as long as the comparison is performed for realistic network architectures with identical conditions, as in (Bredenberg et al., 2021; Payeur et al., 2021; Marschall et al., 2020; Bartunov et al., 2018).

Complexity is multifaceted, and involves features of both stimulus and task (Fig. 2c). Even stimuli with very high dimensional structure can fail to capture critical features of naturalistic stimuli, as evidenced by the wide gap in difficulty involved in constructing convincing models that synthesize images with low-level naturalistic features (orientation, contrast, texture (Portilla and Simoncelli, 2000)) compared to models that capture high-level image features (object identity (Rezende et al., 2014; Goodfellow et al., 2014), semantic content (Ramesh et al., 2021)), which are only just beginning to emerge. Algorithms that scale well with the dimensionality of a stimulus can fail to capture high-level stimulus features: for example, PCA-based image models are unable to capture natural image statistics, and do not result in realistic neural receptive field properties (Olshausen and Field, 1996). For these reasons, it is critical that normative plasticity algorithms be able to scale not just to high-dimensional 'toy' datasets, but also to complex naturalistic data across sensory modalities. This is a major avenue for improvement: for instance, existing plasticity models have great difficulty scaling to naturalistic image datasets (Bartunov et al., 2018).

Similarly, naturalistic task structures are often much more complex than those used for training general machine learning algorithms, let alone models of normative plasticity (Fig. 2c). In natural environments, rewards are often provided after long sequences of complex actions, supervised feedback is sparse, if present at all, and an organism's self preservation often requires navigating both uncertainty and complex multi-agent interactions. Modern reinforcement learning algorithms are only just beginning to make progress with some of these difficulties (Kaelbling et al., 1998; Arjona-Medina et al., 2019; Raposo et al., 2021; Hung et al., 2019; Zhang et al.,

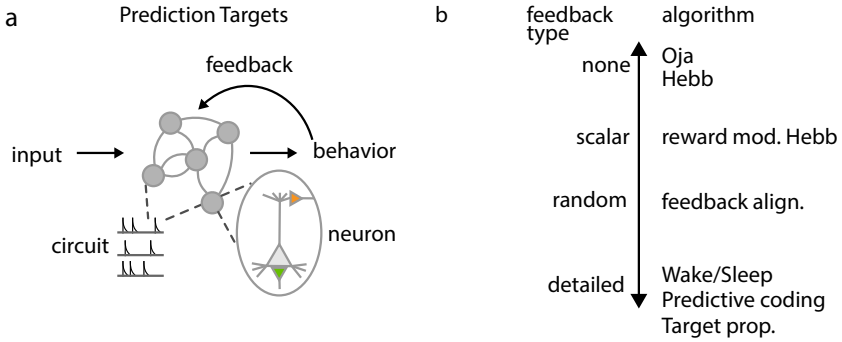

Figure 3: **Testing normative theories. a.** Normative plasticity theories can be assessed through four different experimental lenses centered on individual neurons, circuits of collectively recorded neurons, the training signals delivered to a circuit, and the organism's overall behavior over the course of learning. **b.** Different normative plasticity theories postulate different levels of detail for the feedback signals received by individual neurons.

2021), but as yet there are no normative plasticity models that describe how any of the human capabilities used to solve these problems could be learned through cellular adaptation (for example, model-based planning (Doll et al., 2012)); similarly, none of these capabilities have been shown to be an emergent consequence of a more basic plasticity process.

2.7 Generating testable predictions

Despite the abundance of existing normative theories, very few have been confirmed experimentally, and of those that have received partial confirmation, they are restricted to very specific experimental preparations, for example: fear conditioning in *Aplysia* (Rayport and Schacher, 1986), and reward-based learning in songbird motor systems (Fiete et al., 2007) and in mouse auditory cortex (Froemke et al., 2013; Guo et al., 2019). This relative paucity of validation will not be overcome without a very clear articulation of which features of a normative theory constitute testable predictions, and in what way those predictions disambiguate one theory from its alternatives.

Many existing features of normative theories would be fatal to those theories if proven not to hold in biology. Some examples include: weight symmetry, reward modulation of plasticity, differential roles (and plasticity rules) for apical and basal synapses, and the existence of eligibility traces for temporal credit assignment. However, these individual features, if *proven* to hold, would eliminate alternative theories to highly variable degrees. Most, if not all models could accommodate weight symmetry, several distinct models predict reward modulation of plasticity either through precise credit assignment or global neurotransmitter delivery (Murray, 2019; Williams, 1992; Bellec et al., 2020; Roth et al., 2018), and several distinct supervised and unsupervised models predict different types of signaling and plasticity at apical and basal

synapses on pyramidal neurons (Urbanczik and Senn, 2014; Payeur et al., 2021; Bredenberg et al., 2021; Körding and König, 2001; Schiess et al., 2016; Sacramento et al., 2017; Guerguiev et al., 2017; Richards and Lillicrap, 2019), while nearly all models capable of temporal credit assignment assume some form of synaptic eligibility trace (Bellec et al., 2020; Marschall et al., 2020; Murray, 2019; Miconi, 2017; Roth et al., 2018). It is intuitively clear that for any given normative theory of synaptic plasticity, there exist an infinite number of infinitesimal perturbations to that theory that would be impossible to disambiguate experimentally. Further, there are many features of normative theories that would be fatal if proven not to hold, but are completely unclear how to test experimentally.

The most useful predictions are those that are fatal to the theory if proven false, are clearly testable, and disambiguate the theory from the greatest number of alternative theories. It may be that a collection of predictions is required to completely isolate one individual normative theory from closely related models, which suggests that articulating where particular models lie within a taxonomy of predictions is the most useful way to narrow down the field of possible models. Testable predictions can be defined in terms of several different experimental lenses, of which we isolate four: experiments examining individual neurons or synapses, populations of neurons, the feedback mechanisms that shape learning in neural circuits, or learning at a behavioral level (Fig. 3a). Accurately distinguishing one mechanism from another will likely require a synthesis of experiments spanning all four lenses.

Individual neurons. Experiments that focus on individual neurons, including paired-pulse stimulation (Markram et al., 1997), mechanistic characterizations of plasticity (Graupner and Brunel, 2010), pharmacological explorations of neuromodulators that induce or modify plasticity (Bear and Singer, 1986; Reynolds and Wickens, 2002; Froemke et al., 2007; Gu and Singer, 1995), and characterization of local dendritic or microcircuit properties mediating plasticity (Froemke et al., 2005; Letzkus et al., 2006; Sjöström and Häusser, 2006) form the bulk of the classical literature underlying phenomenological and mechanistic modeling. These studies characterize what information is locally available at synapses and what can be done with that information, as well as which properties of cells can be altered in an experience-dependent fashion.

Existing normative theories differ in the nature of their predictions for plasticity at individual neurons. Reward-modulated Hebbian theories *require* feedback information be delivered by a neuromodulator like dopamine, serotonin, or acetylcholine (Frémaux and Gerstner, 2016) and that this feedback modulates plasticity at the local synapse by changing the magnitude or sign of plasticity depending on the strength of feedback. In contrast, some unsupervised normative theories require no feedback modulation of plasticity (Pehlevan et al., 2015, 2017), and others argue that detailed feedback information arrives at the apical dendritic arbors of pyramidal neurons to modulate plasticity, which is also partially supported in the hippocampus (Bittner et al., 2015, 2017) and cortex (Larkum et al., 1999; Letzkus et al., 2006; Froemke et al., 2005; Sjöström and Häusser, 2006).

Independent of the exact feedback mechanism, models differ in how temporal asso-

ciations are formed. Algorithms related to REINFORCE assume that local synaptic eligibility traces integrate over time fluctuations in coactivity of the post- and presynaptic neuron local to a synapse. These postulated eligibility traces are stochastic, summing Gaussian fluctuations in activity (Miconi, 2017) that consequently produce temporal profiles similar to Brownian motion. In contrast, methods based on approximations to real-time recurrent learning propose eligibility traces that are deterministic records of coactivity whose time constants are directly connected to the dynamics of the neuron itself (Bellec et al., 2020), while other hybrid approaches predict eligibility traces which are deterministic but are related more to predicted task timescale than the dynamics of the cell (Roth et al., 2018). Though there do exist known cellular processes that naturally track coactivity, like NMDA receptors (Bi and Poo, 1998), and that store traces of this coactivity longitudinally, like CaMKII (Graupner and Brunel, 2010), much work remains to be done to analyze how the properties of these known biophysical quantities relate to the predictions of various normative theories, and whether there are other biological alternatives. Other algorithms have different predictions at a microcircuit, rather than at an individual neuron level. Impression learning, for instance, suggests that a population of inhibitory interneurons could gate the influence of apical and basal dendritic inputs to the activity of pyramidal neurons (Bredenberg et al., 2021), and some forms of predictive coding propose that top-down error signals are partially computed by local inhibitory interneurons. Therefore, to completely distinguish different theories, it may be necessary to analyze the connectivity and plasticity between small groups of different cell types.

In sum, experiments at the level of individual neurons or local microcircuits potentially have a great deal of power to identify whether a particular neural circuit is implementing any of a collection of hypothesized normative models of plasticity. It is an advantage that these methods can identify the adaptive capabilities of individual neurons and synapses, but these methods are also limited in their ability to simultaneously observe the adaptation of many neurons in a circuit. Normativity is inherently concerned with the value of plasticity for perception and behavior, and as we will see in subsequent sections, experiments targeting larger populations of neurons will be necessary to distinguish certain features of these theories.

Neural circuits. Determining how circuits encode environmental information and affect motor actions by an animal cannot be assessed by looking at single neurons, and by extension, analyzing how these properties change over time requires methods that record large groups of neurons, such as 2 photon calcium imaging, multielectrode recordings, fMRI, EEG, and MEG, as well as methods that manipulate large populations, like optogenetic (Rajasethupathy et al., 2016) stimulation. The benefits of these recording techniques for testing normative plasticity models, though less practiced compared to individual neuron studies, are manyfold. One of the challenges for characterizing a circuit with a normative plasticity model is selecting an appropriate objective function. Determining which objective fits best can partly be determined by philosophical considerations (Section 2.1), but empirical validation is a far more rigorous test. For instance, one can establish that explicit reward modifies a neural representation to improve coding of task-relevant variables (Froemke et al.,

2013). Another line of approaches trains neural networks on a battery of objectives, and determines which objective produces the closest correspondence between model neurons and neurons recorded brain in a variety of areas in the ventral (Yamins et al., 2014; Yamins and DiCarlo, 2016) and dorsal (Mineault et al., 2021) visual streams, as well as recently in auditory cortex (Kell et al., 2018) and medial entorhinal cortex (Nayebi et al., 2021). Oftentimes, changes in artificial neural network activity throughout time are sufficient to determine the objective optimized by the network as well as its learning algorithm (Nayebi et al., 2020), an approach which could also potentially be applied to recorded neural activity over learning.

Beyond narrowing down the objective function, recording from populations can establish features of neural learning that normative models must account for. For instance, in biofeedback training settings, animals can selectively control the firing rates of individual neurons to satisfy arbitrary experimental conditions for reward (Fetz, 2007), suggesting the existence of highly flexible credit assignment systems in the brain, which are not constrained by evolutionary predetermination of the function of neural circuits[2]. Further, circuit recordings could in principle test predictions about how neural circuits should function in situations that do not specifically involve learning. For instance, the Wake-Sleep algorithm (Dayan et al., 1995) (Appendix D) proposes that neural circuits should spend extended periods of time (e.g. during dreaming) generating similar activity patterns to those evoked by natural stimulus sequences, whereas impression learning proposes that similar hallucinatory states could be induced by experimentally increasing the influence of apical dendrites on pyramidal neuron activity (Bredenberg et al., 2021). An alternative learning algorithm based on generative adversarial networks proposes that during sleep networks rehearse corrupted versions of recent waking experiences (Deperrois et al., 2021). There is plenty of room for experiments to more clearly map predictions and components of these models onto well documented neural phenomena, such as sleep or potentially replay phenomena (Girardeau et al., 2009; Eschenko et al., 2008). Because circuit recording and manipulation methods often sacrifice temporal resolution (Hong and Lieber, 2019), and have difficulty inferring biophysical properties of individual synapses and cells, these methods are best used in concert with single neuron studies to jointly tease apart the multi-level predictions of various normative models.

Feedback mechanisms. One of the best ways to distinguish normative plasticity algorithms is on the basis of the nature of their feedback mechanisms (Fig. 3b). Though some unsupervised algorithms, like Oja's rule propose that no feedback is necessary to perform meaningful learning, no current normative theories propose any form of supervised or reinforcement learning that does not require *some* form of top-down feedback. However, across these models, the level of precision of feedback varies considerably. The simplest feedback is scalar, conveying reward (Williams, 1992), state fluctuation (Payeur et al., 2021), or context (e.g. saccade (Illing et al., 2021) or attention (Roelfsema and Ooyen, 2005; Pozzi et al., 2020)) information. Beyond this, the space of proposed mechanisms expands considerably: backpropa-

[2]This is a challenge for normative plasticity models that predefine the outputs of the circuit and approximately backpropagate errors from these outputs.

gation approximations like feedback alignment (Lillicrap et al., 2016) and random-feedback online learning (RFLO) (Murray, 2019) propose random feedback between layers of neurons can provide a sufficient learning signal, whereas algorithms based on control theory propose that low-rank or partially random projections carrying supervised error signals are sufficient (Gilra and Gerstner, 2017; Alemi et al., 2018). Other algorithms propose even more detailed feedback, with individual neurons receiving precise, carefully adapted projections carrying learning-related information. These algorithms propose that top-down projections to apical dendrites (Urbanczik and Senn, 2014) or local interneurons neurons (Bastos et al., 2012) perform spatial credit assignment, but the nature of this signal can differ considerably across different algorithms. It could be a supervised target, carrying information about what the neuron state 'should' be to achieve a goal (Guerguiev et al., 2017; Payeur et al., 2021), or it could be a prediction of the future state of the neuron (Bredenberg et al., 2020).

Each of these different possibilities is theoretically testable, if the focus is shifted to the postulated feedback mechanism, instead of the circuit undergoing learning. However, so far the different mechanisms have received only partial support. For example, acetylcholine projections to auditory cortex that modulate perceptual learning (Froemke et al., 2013) display a diversity of responses related to both reward and attention (Hangya et al., 2015), which adapt over the course of learning in concert with auditory cortex (Guo et al., 2019). This suggests that while traditional models of reward-modulated Hebbian plasticity may be correct to a first approximation, a more detailed study of the adaptive capabilities of neuromodulatory centers may be necessary to update the theories.

While a growing number of studies indicate that projections to apical synapses of pyramidal neurons *do* play a role in inducing plasticity, and that these projections themselves are also plastic (i.e. nonrandom) (Bittner et al., 2015, 2017), very little is known about the *nature* of the signal—a critical component for distinguishing several different theories. In the visual system, presentation of unfamiliar images without any form or reward or supervision can modify both apical and basal dendrites throughout time (Gillon et al., 2021), and in the hippocampus, apical input to CA1 pyramidal neurons while animals acclimatize to new spatial environments is sufficient to induce synaptic plasticity (Bittner et al., 2015, 2017). These two examples support a form of *unsupervised* learning, but evidence for supervised or reinforcement learning signals propagated through apical dendritic synapses is currently lacking. Beyond the cerebellar system, where climbing fiber pathways may carry explicit motor error signals used for plasticity (Gao et al., 2012; Bouvier et al., 2018), evidence for detailed supervised feedback is limited. In sum, beyond single neurons, or even populations recorded by traditional techniques, targeted focus on the learning feedback signals received by a population shows promise to rule out algorithms on the basis of their feedback and objective function.

Behavior. In much the same way that psychophysical studies of human or animal responses define constraints on what the brain's perceptual systems are capable of, behavioral studies of learning can do quite a lot to describe the range of phenomena that a model of learning must be able to capture, from operant conditioning (Niv,

2009), to model-based learning (Doll et al., 2012), rapid language learning (Heibeck and Markman, 1987), unsupervised sensory development (Wiesel and Hubel, 1963), or consolidation effects (Stickgold, 2005). Behavioral studies can also outline key limitations in learning, which are perhaps reflective of the brain's learning algorithms, including the brain's failure to perform certain types of adaptation after critical periods of plasticity (Wiesel and Hubel, 1963), and the brain's unexpected inability to learn multi-context motor movements without explicit motor differences across contexts (Sheahan et al., 2016).

These existing experimental results stand as (often unmet) targets for normative theories of plasticity, but in addition, normative theories themselves suggest further studies that may test their predictions. In particular, manipulation of learning mechanisms may have predictable effects on animals' behavior, as seen when acetylcholine receptor blockage in mouse auditory cortex prevented reward-based learning in animals (Guo et al., 2019), and nucleus basalis stimulation during tone perception longitudinally improved animals' discrimination of that tone (Froemke et al., 2013). Other algorithms have as-yet untested predictions for behavior: for instance, experimentally increasing the influence of top-down projections should bias behavior towards commonly-occurring sensory stimuli according to both predictive coding (Rao and Ballard, 1999; Friston, 2010) and impression learning (Bredenberg et al., 2021). For other detailed feedback algorithms (Fig. 3b), manipulating top-down projections may disrupt learning, but would have a much more unstructured deleterious effect on perceptual behavior.

As shown, each experimental lens has its own advantages and disadvantages. Single-neuron studies are excellent for identifying the locally-available variables that affect plasticity, circuit-level studies can help narrow down the objectives that shape neural responses and identify traces of offline learning, studies of feedback mechanisms can distinguish between different algorithms that postulate different degrees of precision in their feedback and in complexity of the teaching signal, and studies of behavior can place boundaries on what can be learned, as well as serve as a readout for manipulations of the mechanisms underlying learning. Each focus alone is insufficient to distinguish between all existing normative models, but in concert they show promise for identifying the neural substrates of adaptation.

3 Conclusions

Normative plasticity models are compelling because of their potential to connect our brains' capacity for adaptation to their constituent synaptic modifications. Generating good theories is a critical part of the scientific process, but finding ways to close the loop by testing key predictions of new normative models has proved extraordinarily difficult: in this perspective we have illustrated the sources of this difficulty.

Algorithm	Dec. Loss	Local	Arch.	Time	Online	Scalable
Backpropagation (Wer74)	U/S/R (Wil92)	✗	✓ (Lee16)	✓ (Wer90)	✗	✓
REINFORCE (Wil92)	U/S/R	✓	✓	✓ (Mic17)	✓	✗ (Wer03)
Oja (Oja82)	U	✓	✗	✗	✓	✓
Predictive Coding (Rao99)	U/S (Whi17)	✓	✗	✓ (Fri09)	✓	✓
Wake-Sleep (Day95)	U	✓	✓ (Day96)	✓ (Day96)	✓ (Bre21)	✓
Approx. Backprop. (Lil16) (Akr19)	U/S*	✓	✓ (Bel20)	✓ (Mur19) (Bel20)	✓ (Mur19) (Bel20)	✓
Equilibrium Prop. (Sce17)	U/S	✓	✗	✗	✓ (Ern20)	✓ (Lab21)
Target Prop. (Ben14)	U/S	✓	✓	✓ (Man20)	✗	✓ (Lee15)

Table 1: **Summarizing progress on the desiderata.** A ✓ indicates that an algorithm has been demonstrated to satisfy a particular desideratum in at least one study, whereas an ✗ indicates that it has not been demonstrated. If the demonstrating study is an improvement on the seminal work or is a new model, we provide a citation; reference numbering used for brevity: Asterisks (*) indicate that results have only been shown by simulation, and lack mathematical support. U, S, and R indicate whether a given algorithm supports unsupervised, supervised, or reinforcement learning, respectively.

The core of a normative plasticity model is its plasticity rule, which dictates how a model synapse modifies its strength. To be a normative model—to explain why the plasticity mechanism is important for the organism—there must be a concrete demonstration that this plasticity rule supports adaptation critical for system-wide goals like processing sensory signals or obtaining rewards (Section 2.1). However, this system-wide goal must be achieved using only *local* information (Section 2.2). These two needs of a normative plasticity model are the fundamental source of tension: it is very difficult to demonstrate that a proposed plasticity rule is both local *and* optimizes a system-wide objective (Appendix B). Insufficient or partial resolution of this fundamental tension produces normative models that struggle to map accurately onto neural hardware (Section 2.3) or handle complex temporal stimuli and tasks online (Sections 2.4-2.6). To provide a case study of how our desiderata come to be satisfied (or not) in practice, we have included tutorials for both REINFORCE and the Wake-Sleep algorithm in Appendices C and D. These tutorials are by no means a complete introduction to the field, but will hopefully serve as a solid foothold for analyzing modern normative plasticity models.

Even satisfying the aforementioned desiderata, much work remains to delineate which tests would most clearly distinguish a normative model from its alternatives in a biological system. In this review, we have organized emerging theories

according to how they satisfy and improve upon our desiderata (Table 1), as well as by how they can be tested (Section 2.7), with the view that this organization will provide avenues for both experimental and theoretical neuroscientists to bring normative plasticity models closer to biology. Even if existing algorithms prove not to be implemented exactly in the brain, they undoubtedly provide key insights into how local synaptic modifications can produce valuable improvements in both behavior and perception for an organism. It seems sensible to use these algorithms as a springboard to produce more biologically realistic and powerful theories.

Beyond improving normative theories with respect to our desiderata, there are several incredible opportunities for actually testing their implementation in biology (Section 2.7). Most current theoretical studies of reward-modulated Hebbian plasticity focus on dopamine-modulated motor learning in monkeys and songbirds (Fiete et al., 2007; Legenstein et al., 2010), but there are *many* neuromodulatory systems that have been linked to learning in experiments, including serotonin-modulated fear conditioning in the amygdala (Lesch and Waider, 2012), as well as acetylcholine-modulated reward learning and oxytocin-modulated social learning in mouse auditory cortex (Guo et al., 2019; Froemke et al., 2013). Further, several experimental preparations examine the relationship between pyramidal neurons' apical and basal dendritic activity and plasticity, in both the hippocampus (Bittner et al., 2015, 2017) and visual cortex (Gillon et al., 2021; Froemke et al., 2005; Letzkus et al., 2006; Sjöström and Häusser, 2006). These could test at the level of individual neurons, circuits, behavior, and the feedback mechanisms that support plasticity, which of the many alternative normative theories underlie animals' learning.

As the diversity of aforementioned experimental preparations suggests, there are increasingly strong arguments for several fundamentally different plasticity algorithms instantiated in different areas of the brain and across different organisms, subserving different functions. It is quite likely that many plasticity mechanisms work in concert to produce learning as it manifests in our perception and behavior. It is our belief that well-articulated normative theories can serve as the building blocks of a conceptual framework that tames this diversity and allows us to understand the brain's tremendous capacity for adaptation.

4 Acknowledgements and Funding

We would like to thank Blake Richards, Eero Simoncelli, Owen Marschall, Benjamin Lyo, Elliott Capek, Olivier Codol, and Yuhe Fan for their helpful feedback on this manuscript. CS is supported by NIMH Award 1R01MH125571-01, NIH Award R01NS127122, by the National Science Foundation under NSF Award No. 1922658 and a Google faculty award.

References

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive science*, 9(1):147–169.

Akrout, M., Wilson, C., Humphreys, P. C., Lillicrap, T., and Tweed, D. (2019). Using weight mirrors to improve feedback alignment. *arXiv preprint arXiv:1904.05391*.

Alemi, A., Machens, C., Deneve, S., and Slotine, J.-J. (2018). Learning nonlinear dynamics in efficient, balanced spiking networks using local plasticity rules. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Amari, S. I. S.-i. and Nakahara, H. (1999). Convergence of the wake-sleep algorithm. In *Advances in Neural Information Processing Systems 11: Proceedings of the 1998 Conference*, volume 11, page 239. MIT Press.

Arjona-Medina, J. A., Gillhofer, M., Widrich, M., Unterthiner, T., Brandstetter, J., and Hochreiter, S. (2019). Rudder: Return decomposition for delayed rewards. *Advances in Neural Information Processing Systems*, 32.

Atick, J. J. and Redlich, A. N. (1990). Towards a theory of early visual processing. *Neural computation*, 2(3):308–320.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological review*, 61(3):183.

Bartunov, S., Santoro, A., Richards, B., Marris, L., Hinton, G. E., and Lillicrap, T. (2018). Assessing the scalability of biologically-motivated deep learning algorithms and architectures. *Advances in neural information processing systems*, 31.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711.

Bear, M. F. and Singer, W. (1986). Modulation of visual cortical plasticity by acetylcholine and noradrenaline. *Nature*, 320(6058):172–176.

Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., and Maass, W. (2020). A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature communications*, 11(1):1–15.

Bengio, Y. (2014). How auto-encoders could provide credit assignment in deep networks via target propagation. *arXiv preprint arXiv:1407.7906*.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Benna, M. K. and Fusi, S. (2016). Computational principles of synaptic memory consolidation. *Nature neuroscience*, 19(12):1697–1706.

Bi, G.-q. and Poo, M.-m. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of neuroscience*, 18(24):10464–10472.

Bittner, K. C., Grienberger, C., Vaidya, S. P., Milstein, A. D., Macklin, J. J., Suh, J., Tonegawa, S., and Magee, J. C. (2015). Conjunctive input processing drives feature selectivity in hippocampal CA1 neurons. *Nature neuroscience*, 18(8):1133–1142.

Bittner, K. C., Milstein, A. D., Grienberger, C., Romani, S., and Magee, J. C. (2017). Behavioral time scale synaptic plasticity underlies CA1 place fields. *Science*, 357(6355):1033–1036.

Bliss, T. V. and Collingridge, G. L. (1993). A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*, 361(6407):31–39.

Bönstrup, M., Iturrate, I., Thompson, R., Cruciani, G., Censor, N., and Cohen, L. G. (2019). A rapid form of offline consolidation in skill learning. *Current Biology*, 29(8):1346–1351.

Bouvier, G., Aljadeff, J., Clopath, C., Bimbard, C., Ranft, J., Blot, A., Nadal, J.-P., Brunel, N., Hakim, V., and Barbour, B. (2018). Cerebellar learning using perturbations. *Elife*, 7:e31599.

Bowers, J. S. and Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological bulletin*, 138(3):389.

Bredenberg, C., Lyo, B. S. H., Simoncelli, E. P., and Savin, C. (2021). Impression learning: Online representation learning with synaptic plasticity. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Bredenberg, C., Simoncelli, E., and Savin, C. (2020). Learning efficient task-dependent representations with synaptic plasticity. *Advances in Neural Information Processing Systems*, 33.

Brendel, W., Bourdoukan, R., Vertechi, P., Machens, C. K., and Denéve, S. (2020). Learning to represent signals spike by spike. *PLoS computational biology*, 16(3):e1007692.

Brody, C. D., Hernández, A., Zainos, A., and Romo, R. (2003). Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. *Cerebral cortex*, 13(11):1196–1207.

Calabresi, P., Picconi, B., Tozzi, A., and Di Filippo, M. (2007). Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends in neurosciences*, 30(5):211–219.

Cartwright, N. and McMullin, E. (1984). How the laws of physics lie.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Clark, A. and Toribio, J. (1994). Doing without representing? *Synthese*, 101(3):401–431.

Clopath, C., Ziegler, L., Vasilaki, E., Büsing, L., and Gerstner, W. (2008). Tag-trigger-consolidation: a model of early and late long-term-potentiation and depression. *PLoS computational biology*, 4(12):e1000248.

Compte, A., Brunel, N., Goldman-Rakic, P. S., and Wang, X.-J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral cortex*, 10(9):910–923.

Cornford, J., Kalajdzievski, D., Leite, M., Lamarquette, A., Kullmann, D. M., and Richards, B. (2021). Learning to live with Dale's principle: ANNs with separate excitatory and inhibitory units. *bioRxiv*, pages 2020–11.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford University Press.

Dan, Y. and Poo, M.-m. (2004). Spike timing-dependent plasticity of neural circuits. *Neuron*, 44(1):23–30.

Dayan, P. and Hinton, G. E. (1996). Varieties of Helmholtz machine. *Neural Networks*, 9(8):1385–1403.

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The Helmholtz machine. *Neural computation*, 7(5):889–904.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Deperrois, N., Petrovici, M. A., Senn, W., and Jordan, J. (2021). Memory semantization through perturbed and adversarial dreaming. *arXiv preprint arXiv:2109.04261*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Doll, B. B., Simon, D. A., and Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current opinion in neurobiology*, 22(6):1075–1081.

Ernoult, M., Grollier, J., Querlioz, D., Bengio, Y., and Scellier, B. (2020). Equilibrium propagation with continual weight updates. *arXiv preprint arXiv:2005.04168*.

Eschenko, O., Ramadan, W., Mölle, M., Born, J., and Sara, S. J. (2008). Sustained increase in hippocampal sharp-wave ripple activity during slow-wave sleep after learning. *Learning & memory*, 15(4):222–228.

Faisal, A. A., Selen, L. P., and Wolpert, D. M. (2008). Noise in the nervous system. *Nature reviews neuroscience*, 9(4):292–303.

Fetz, E. E. (2007). Volitional control of neural activity: implications for brain–computer interfaces. *The Journal of physiology*, 579(3):571–579.

Feulner, B. and Clopath, C. (2021). Neural manifold under plasticity in a goal driven learning behaviour. *PLoS computational biology*, 17(2):e1008621.

Fiete, I. R. (2004). *Learning and coding in biological neural networks*. Harvard University.

Fiete, I. R., Fee, M. S., and Seung, H. S. (2007). Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances. *Journal of neurophysiology*, 98(4):2038–2057.

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in cognitive sciences*, 14(3):119–130.

Frémaux, N. and Gerstner, W. (2016). Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Frontiers in neural circuits*, 9:85.

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

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138.

Friston, K. and Kiebel, S. (2009). Cortical circuits for perceptual inference. *Neural Networks*, 22(8):1093–1104.

Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature neuroscience*, 6(11):1216–1223.

Froemke, R. C., Carcea, I., Barker, A. J., Yuan, K., Seybold, B. A., Martins, A. R. O., Zaika, N., Bernstein, H., Wachs, M., Levis, P. A., et al. (2013). Long-term modification of cortical synapses improves sensory perception. *Nature neuroscience*, 16(1):79–88.

Froemke, R. C., Merzenich, M. M., and Schreiner, C. E. (2007). A synaptic memory trace for cortical receptive field plasticity. *Nature*, 450(7168):425–429.

Froemke, R. C., Poo, M.-m., and Dan, Y. (2005). Spike-timing-dependent synaptic plasticity depends on dendritic location. *Nature*, 434(7030):221–225.

Fukushima, K. and Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer.

Fusi, S., Drew, P. J., and Abbott, L. F. (2005). Cascade models of synaptically stored memories. *Neuron*, 45(4):599–611.

Ganguli, S., Huh, D., and Sompolinsky, H. (2008). Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences*, 105(48):18970–18975.

Gao, Z., Van Beugen, B. J., and De Zeeuw, C. I. (2012). Distributed synergistic plasticity and cerebellar learning. *Nature Reviews Neuroscience*, 13(9):619–635.

Gerstner, W. and Kistler, W. M. (2002). *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press.

Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D., and Brea, J. (2018). Eligibility traces and plasticity on behavioral time scales: experimental support of neoHebbian three-factor learning rules. *Frontiers in neural circuits*, 12:53.

Gillon, C. J., Pina, J. E., Lecoq, J. A., Ahmed, R., Billeh, Y., Caldejon, S., Groblewski, P., Henley, T. M., Lee, E., Luviano, J., et al. (2021). Learning from unexpected events in the neocortical microcircuit. *bioRxiv*.

Gilra, A. and Gerstner, W. (2017). Predicting non-linear dynamics by stable local learning in a recurrent spiking neural network. *Elife*, 6:e28295.

Girardeau, G., Benchenane, K., Wiener, S. I., Buzsáki, G., and Zugaro, M. B. (2009). Selective suppression of hippocampal ripples impairs spatial memory. *Nature neuroscience*, 12(10):1222–1223.

Gold, J. I. and Shadlen, M. N. (2007). The neural basis of decision making. *Annu. Rev. Neurosci.*, 30:535–574.

Golkar, S., Tesileanu, T., Bahroun, Y., Sengupta, A. M., and Chklovskii, D. B. (2022). Constrained predictive coding as a biologically plausible model of the cortical hierarchy. *arXiv preprint arXiv:2210.15752*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Graupner, M. and Brunel, N. (2010). Mechanisms of induction and maintenance of spike-timing dependent plasticity in biophysical synapse models. *Frontiers in computational neuroscience*, 4:136.

Grienberger, C. and Magee, J. C. (2022). Entorhinal cortex directs learning-related changes in CA1 representations. *Nature*, pages 1–9.

Gu, Q. and Singer, W. (1995). Involvement of serotonin in developmental plasticity of kitten visual cortex. *European Journal of Neuroscience*, 7(6):1146–1153.

Guerguiev, J., Lillicrap, T. P., and Richards, B. A. (2017). Towards deep learning with segregated dendrites. *Elife*, 6:e22901.

Gulati, T., Guo, L., Ramanathan, D. S., Bodepudi, A., and Ganguly, K. (2017). Neural reactivations during sleep determine network credit assignment. *Nature neuroscience*, 20(9):1277–1284.

Guo, W., Robert, B., and Polley, D. B. (2019). The cholinergic basal forebrain links auditory stimuli with delayed reinforcement to support learning. *Neuron*, 103(6):1164–1177.

Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. (2019). Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.

Hangya, B., Ranade, S. P., Lorenc, M., and Kepecs, A. (2015). Central cholinergic neurons are rapidly recruited by reinforcement feedback. *Cell*, 162(5):1155–1168.

Hebb, D. O. (1949). *The organisation of behaviour: a neuropsychological theory*. Science Editions New York.

Heess, N., TB, D., Sriram, S., Lemmon, J., Merel, J., Wayne, G., Tassa, Y., Erez, T., Wang, Z., Eslami, S., et al. (2017). Emergence of locomotion behaviours in rich environments. *arXiv preprint arXiv:1707.02286*.

Heibeck, T. H. and Markman, E. M. (1987). Word learning in children: An examination of fast mapping. *Child development*, pages 1021–1034.

Hennequin, G., Vogels, T. P., and Gerstner, W. (2012). Non-normal amplification in random balanced neuronal networks. *Physical Review E*, 86(1):011909.

Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hong, G. and Lieber, C. M. (2019). Novel electrode technologies for neural recordings. *Nature Reviews Neuroscience*, 20(6):330–345.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558.

Hung, C.-C., Lillicrap, T., Abramson, J., Wu, Y., Mirza, M., Carnevale, F., Ahuja, A., and Wayne, G. (2019). Optimizing agent behavior over long time scales by transporting value. *Nature communications*, 10(1):1–12.

Illing, B., Ventura, J., Bellec, G., and Gerstner, W. (2021). Local plasticity rules can learn deep representations using self-supervised contrastive predictions. *Advances in Neural Information Processing Systems*, 34.

Jabri, M. and Flower, B. (1992). Weight perturbation: An optimal architecture and learning technique for analog VLSI feedforward and recurrent multilayer networks. *IEEE Transactions on Neural Networks*, 3(1):154–157.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134.

Kappel, D., Nessler, B., and Maass, W. (2014). STDP installs in winner-take-all circuits an online approximation to hidden markov model learning. *PLoS computational biology*, 10(3):e1003511.

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., and McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes.

Kirby, K. G. (2006). A tutorial on Helmholtz machines. *Department of Computer Science, Northern Kentucky University*.

Körding, K. P. and König, P. (2001). Supervised and unsupervised learning with two sites of synaptic integration. *Journal of computational neuroscience*, 11(3):207–215.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Laborieux, A., Ernoult, M., Scellier, B., Bengio, Y., Grollier, J., and Querlioz, D. (2021). Scaling equilibrium propagation to deep convnets by drastically reducing its gradient estimator bias. *Frontiers in neuroscience*, 15:129.

Larkum, M. E., Zhu, J. J., and Sakmann, B. (1999). A new cellular mechanism for coupling inputs arriving at different cortical layers. *Nature*, 398(6725):338–341.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2.

Lee, D.-H., Zhang, S., Fischer, A., and Bengio, Y. (2015). Difference target propagation. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 498–515. Springer.

Lee, J. H., Delbruck, T., and Pfeiffer, M. (2016). Training deep spiking neural networks using backpropagation. *Frontiers in neuroscience*, 10:508.

Legenstein, R., Chase, S. M., Schwartz, A. B., and Maass, W. (2010). A reward-modulated Hebbian learning rule can explain experimentally observed network reorganization in a brain control task. *Journal of Neuroscience*, 30(25):8400–8410.

Lesch, K.-P. and Waider, J. (2012). Serotonin in the modulation of neural plasticity and networks: implications for neurodevelopmental disorders. *Neuron*, 76(1):175–191.

Letzkus, J. J., Kampa, B. M., and Stuart, G. J. (2006). Learning rules for spike timing-dependent plasticity depend on dendritic synapse location. *Journal of Neuroscience*, 26(41):10420–10429.

Levelt, C. N. and Hübener, M. (2012). Critical-period plasticity in the visual cortex. *Annual review of neuroscience*, 35:309–330.

Levenstein, D., Alvarez, V. A., Amarasingham, A., Azab, H., Gerkin, R. C., Hasenstaub, A., Iyer, R., Jolivet, R. B., Marzen, S., Monaco, J. D., et al. (2020). On the role of theory and modeling in neuroscience. *arXiv preprint arXiv:2003.13825*.

Lillicrap, T. P., Cownden, D., Tweed, D. B., and Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications*, 7(1):1–10.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., and Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346.

Liu, Y., Mattar, M. G., Behrens, T. E., Daw, N. D., and Dolan, R. J. (2021). Experience replay is associated with efficient nonlocal learning. *Science*, 372(6544):eabf1357.

Magee, J. C. and Johnston, D. (1997). A synaptically controlled, associative signal for Hebbian plasticity in hippocampal neurons. *Science*, 275(5297):209–213.

Manchev, N. and Spratling, M. W. (2020). Target propagation in recurrent neural networks. *J. Mach. Learn. Res.*, 21(7):1–33.

Mante, V., Sussillo, D., Shenoy, K. V., and Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84.

Markram, H., Lübke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. *Science*, 275(5297):213–215.

Marlin, B. J., Mitre, M., D'amour, J. A., Chao, M. V., and Froemke, R. C. (2015). Oxytocin enables maternal behaviour by balancing cortical inhibition. *Nature*, 520(7548):499–504.

Marschall, O., Cho, K., and Savin, C. (2020). A unified framework of online learning algorithms for training recurrent neural networks. *Journal of machine learning research*.

Martin, S. J., Grimwood, P. D., and Morris, R. G. (2000). Synaptic plasticity and memory: an evaluation of the hypothesis. *Annual review of neuroscience*, 23(1):649–711.

Martins, A. R. O. and Froemke, R. C. (2015). Coordinated forms of noradrenergic plasticity in the locus coeruleus and primary auditory cortex. *Nature neuroscience*, 18(10):1483–1492.

Meulemans, A., Carzaniga, F. S., Suykens, J. A., Sacramento, J., and Grewe, B. F. (2020). A theoretical framework for target propagation. *arXiv preprint arXiv:2006.14331*.

Meulemans, A., Zucchet, N., Kobayashi, S., Von Oswald, J., and Sacramento, J. (2022). The least-control principle for local learning at equilibrium. *Advances in Neural Information Processing Systems*, 35:33603–33617.

Miconi, T. (2017). Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *Elife*, 6:e20899.

Mineault, P., Bakhtiari, S., Richards, B., and Pack, C. (2021). Your head is there to move you around: Goal-driven models of the primate dorsal pathway. *Advances in Neural Information Processing Systems*, 34.

Mnih, A. and Gregor, K. (2014). Neural variational inference and learning in belief networks. In *International Conference on Machine Learning*, pages 1791–1799. PMLR.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Murphy, T. H. and Corbett, D. (2009). Plasticity during stroke recovery: from synapse to behaviour. *Nature reviews neuroscience*, 10(12):861–872.

Murray, J. M. (2019). Local online learning in recurrent networks with random feedback. *ELife*, 8:e43299.

Nayebi, A., Attinger, A., Campbell, M., Hardcastle, K., Low, I., Mallory, C., Mel, G., Sorscher, B., Williams, A., Ganguli, S., et al. (2021). Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks. *Advances in Neural Information Processing Systems*, 34.

Nayebi, A., Srivastava, S., Ganguli, S., and Yamins, D. L. (2020). Identifying learning rules from neural network observables. *arXiv preprint arXiv:2010.11765*.

Neftci, E. O., Mostafa, H., and Zenke, F. (2019). Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63.

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

Obeid, D., Ramambason, H., and Pehlevan, C. (2019). Structured and deep similarity matching via structured and deep Hebbian networks. *arXiv preprint arXiv:1910.04958*.

O'Donohue, T. L., Millington, W. R., Handelmann, G. E., Contreras, P. C., and Chronwall, B. M. (1985). On the 50th anniversary of Dale's law: multiple neurotransmitter neurons. *Trends in Pharmacological Sciences*, 6:305–308.

Ohl, F. W. and Scheich, H. (2005). Learning-induced plasticity in animal and human auditory cortex. *Current opinion in neurobiology*, 15(4):470–477.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Otani, S., Daniel, H., Roisin, M.-P., and Crepel, F. (2003). Dopaminergic modulation of long-term synaptic plasticity in rat prefrontal neurons. *Cerebral cortex*, 13(11):1251–1256.

Pavlides, C. and Winson, J. (1989). Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes. *Journal of neuroscience*, 9(8):2907–2918.

Payeur, A., Guerguiev, J., Zenke, F., Richards, B. A., and Naud, R. (2021). Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits. *Nature neuroscience*, pages 1–10.

Pehlevan, C., Hu, T., and Chklovskii, D. B. (2015). A Hebbian/anti-Hebbian neural network for linear subspace learning: A derivation from multidimensional scaling of streaming data. *Neural computation*, 27(7):1461–1495.

Pehlevan, C., Sengupta, A. M., and Chklovskii, D. B. (2017). Why do similarity matching objectives lead to Hebbian/anti-Hebbian networks? *Neural computation*, 30(1):84–124.

Portilla, J. and Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70.

Pozzi, I., Bohte, S., and Roelfsema, P. (2020). Attention-gated brain propagation: How the brain can implement reward-based error backpropagation. *Advances in Neural Information Processing Systems*, 33.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Rajasethupathy, P., Ferenczi, E., and Deisseroth, K. (2016). Targeting neural circuits. *Cell*, 165(3):524–534.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR.

Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87.

Raposo, D., Ritter, S., Santoro, A., Wayne, G., Weber, T., Botvinick, M., van Hasselt, H., and Song, F. (2021). Synthetic returns for long-term credit assignment. *arXiv preprint arXiv:2102.12425*.

Rasmusson, D. (2000). The role of acetylcholine in cortical synaptic plasticity. *Behavioural brain research*, 115(2):205–218.

Rayport, S. G. and Schacher, S. (1986). Synaptic plasticity in vitro: cell culture of identified Aplysia neurons mediating short-term habituation and sensitization. *Journal of Neuroscience*, 6(3):759–763.

Reynolds, J. N. and Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural networks*, 15(4-6):507–521.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR.

Richards, B. A. and Lillicrap, T. P. (2019). Dendritic solutions to the credit assignment problem. *Current opinion in neurobiology*, 54:28–36.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., et al. (2019). A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770.

Roelfsema, P. R. and Ooyen, A. v. (2005). Attention-gated reinforcement learning of internal representations for classification. *Neural computation*, 17(10):2176–2214.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Roth, C., Kanitscheider, I., and Fiete, I. (2018). Kernel rnn learning (kernl). In *International Conference on Learning Representations*.

Roweis, S. and Ghahramani, Z. (1999). A unifying review of linear gaussian models. *Neural computation*, 11(2):305–345.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Rust, N. C. and DiCarlo, J. J. (2010). Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area v4 to it. *Journal of Neuroscience*, 30(39):12978–12995.

Sacramento, J., Costa, R. P., Bengio, Y., and Senn, W. (2017). Dendritic error backpropagation in deep cortical microcircuits. *arXiv preprint arXiv:1801.00062*.

Savin, C., Joshi, P., and Triesch, J. (2010). Independent component analysis in spiking neurons. *PLoS computational biology*, 6(4):e1000757.

Scellier, B. and Bengio, Y. (2017). Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in computational neuroscience*, 11:24.

Schiess, M., Urbanczik, R., and Senn, W. (2016). Somato-dendritic synaptic plasticity and error-backpropagation in active dendrites. *PLoS computational biology*, 12(2):e1004638.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599.

Shadlen, M. N. and Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of neurophysiology*, 86(4):1916–1936.

Sheahan, H. R., Franklin, D. W., and Wolpert, D. M. (2016). Motor planning, not execution, separates motor memories. *Neuron*, 92(4):773–779.

Sheffield, M. E., Adoff, M. D., and Dombeck, D. A. (2017). Increased prevalence of calcium transients across the dendritic arbor during place field formation. *Neuron*, 96(2):490–504.

Shinoe, T., Matsui, M., Taketo, M. M., and Manabe, T. (2005). Modulation of synaptic plasticity by physiological activation of M1 muscarinic acetylcholine receptors in the mouse hippocampus. *Journal of Neuroscience*, 25(48):11194–11200.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359.

Simoncelli, E. P. (2003). Vision and the statistics of the visual environment. *Current opinion in neurobiology*, 13(2):144–149.

Simoncelli, E. P. and Heeger, D. J. (1998). A model of neuronal responses in visual area mt. *Vision research*, 38(5):743–761.

Simoncelli, E. P. and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216.

Sjöström, J., Gerstner, W., et al. (2010). Spike-timing dependent plasticity. *Spike-timing dependent plasticity*, 35(0):0–0.

Sjöström, P. J. and Häusser, M. (2006). A cooperative switch determines the sign of synaptic plasticity in distal dendrites of neocortical pyramidal neurons. *Neuron*, 51(2):227–238.

Sohn, H., Narain, D., Meirhaeghe, N., and Jazayeri, M. (2019). Bayesian computation through cortical latent dynamics. *Neuron*, 103(5):934–947.

Sompolinsky, H. and Kanter, I. (1986). Temporal association in asymmetric neural networks. *Physical review letters*, 57(22):2861.

Stickgold, R. (2005). Sleep-dependent memory consolidation. *Nature*, 437(7063):1272–1278.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tishby, N., Pereira, F. C., and Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.

Urbanczik, R. and Senn, W. (2014). Learning by the dendritic prediction of somatic spiking. *Neuron*, 81(3):521–528.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., and Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12).

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354.

Vogels, T. P., Sprekeler, H., Zenke, F., Clopath, C., and Gerstner, W. (2011). Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science*, 334(6062):1569–1573.

Werbos, P. (1974). Beyond regression: New tools for prediction and analysis in the behavioral sciences. *Ph. D. dissertation, Harvard University*.

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

Werfel, J., Xie, X., and Seung, H. S. (2003). Learning curves for stochastic gradient descent in linear feedforward networks. In *NIPS*, pages 1197–1204. Citeseer.

Whittington, J. C. and Bogacz, R. (2017). An approximation of the error backpropagation algorithm in a predictive coding network with local Hebbian synaptic plasticity. *Neural computation*, 29(5):1229–1262.

Wiesel, T. N. and Hubel, D. H. (1963). Single-cell responses in striate cortex of kittens deprived of vision in one eye. *Journal of neurophysiology*, 26(6):1003–1017.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.

Williams, R. J. and Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Wong, K.-F. and Wang, X.-J. (2006). A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328.

Xiao, Z.-C., Lin, K. K., and Young, L.-S. (2021). A data-informed mean-field approach to mapping of cortical parameter landscapes. *PLOS Computational Biology*, 17(12):e1009718.

Xie, X. and Seung, H. S. (2003). Equivalence of backpropagation and contrastive Hebbian learning in a layered network. *Neural computation*, 15(2):441–454.

Yamins, D. L. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624.

Zhang, K., Yang, Z., and Başar, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384.

Zhang, L. I., Tao, H. W., Holt, C. E., Harris, W. A., and Poo, M.-m. (1998). A critical window for cooperation and competition among developing retinotectal synapses. *Nature*, 395(6697):37–44.

Ziemba, C. M., Freeman, J., Movshon, J. A., and Simoncelli, E. P. (2016). Selectivity and tolerance for visual texture in macaque V2. *Proceedings of the National Academy of Sciences*, 113(22):E3140–E3149.

Zigmond, M. J., Abercrombie, E. D., Berger, T. W., Grace, A. A., and Stricker, E. M. (1990). Compensations after lesions of central dopaminergic neurons: some clinical and basic implications. *Trends in neurosciences*, 13(7):290–296.

A The unidentifiability of an objective

In this section we illustrate why the choice of objective function for a normative plasticity model is never uniquely determined by data. We will consider two situations: the system has already settled to its optimal setting of its weights, $\mathbf{W}^*$, and in the second we are able to observe the system's plasticity update $\Delta\mathbf{W}$.

A.1 Unidentifiability based on an optimum

Suppose that some setting of synaptic weights $\mathbf{W}^*$ minimizes an objective function $\mathcal{L}$, i.e. $\mathcal{L}(\mathbf{W}^*) \leq \mathcal{L}(\mathbf{W})\ \forall\mathbf{W}$. We might be tempted to argue that because $\mathbf{W}^*$ minimizes $\mathcal{L}$, $\mathcal{L}$ must be *the* objective that the system is minimizing. However, there are an infinite variety of alternative objectives that share the same minimum. To see this, take a new objective $\tilde{\mathcal{L}} = \sigma\left(\mathcal{L}(\mathbf{W})\right)$ for any differentiable, monotonically increasing function $\sigma(\cdot)$. Then we have:

$$\mathcal{L}(\mathbf{W}^*) \leq \mathcal{L}(\mathbf{W})\ \forall\mathbf{W} \tag{4}$$

$$\Rightarrow \sigma\left(\mathcal{L}(\mathbf{W}^*)\right) \leq \sigma\left(\mathcal{L}(\mathbf{W})\right)\ \forall\mathbf{W} \tag{5}$$

$$\Rightarrow \tilde{\mathcal{L}}(\mathbf{W}^*) \leq \tilde{\mathcal{L}}(\mathbf{W})\ \forall\mathbf{W}, \tag{6}$$

where the second equality follows from the order preservation property of $\sigma(\cdot)$. This means that $\mathbf{W}^*$ also minimizes $\tilde{\mathcal{L}}$, i.e. we will be unable to arbitrate between whether the system is 'attempting' to minimize $\tilde{\mathcal{L}}$ or $\mathcal{L}$ on the basis of the optimized network state given by $\mathbf{W}^*$.

A.2 Unidentifiability based on an update rule

Suppose instead that we were able to observe the adaptive plasticity mechanism of a system, and were able to verify that it really does decrease an objective function $\mathcal{L}$, i.e. by Eq. 3,

$$\frac{d\mathcal{L}}{d\mathbf{W}}(\mathbf{W})^T\Delta\mathbf{W} \leq 0\ \forall\mathbf{W}. \tag{7}$$

We might now be tempted to argue that, by observing the plasticity rule itself, $\Delta\mathbf{W}$, we will be more able to assert that the system, by virtue of consistently decreasing $\mathcal{L}$, is 'attempting' to minimize $\mathcal{L}$. However, the *exact same* family of alternative objectives will also be minimized ($\tilde{\mathcal{L}} = \sigma\left(\mathcal{L}(\mathbf{W})\right)$ for any differentiable, monotonically increasing function $\sigma(\cdot)$). To see this, we observe:

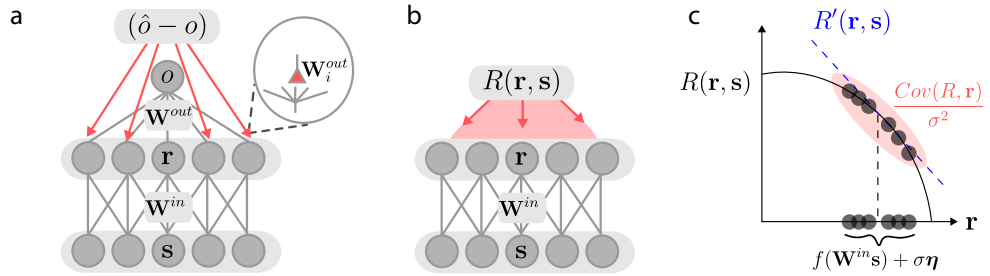

Figure S1: **Weight transport and REINFORCE. a.** Traditional gradient descent propagates a credit assignment signal $(\hat{o} - o)\mathbf{W}_i^{out}$ to each neuron $\mathbf{r}_i$. How this pathway could have access to $\mathbf{W}_i^{out}$ is unclear: this is the 'weight transport' problem. **b.** REINFORCE resolves the weight transport problem by projecting a scalar reward signal $R(\mathbf{r}, \mathbf{s})$ to all synapses. **c.** By correlating this reward with fluctuations in neural activity, neurons can approximate the true gradient.

$$\frac{d\mathcal{L}}{d\mathbf{W}}(\mathbf{W})^T \Delta\mathbf{W} \leq 0 \ \forall \mathbf{W} \tag{8}$$

$$\Rightarrow \frac{d\sigma(\mathcal{L}(\mathbf{W}))}{d\mathcal{L}(\mathbf{W})} \frac{d\mathcal{L}}{d\mathbf{W}}(\mathbf{W})^T \Delta\mathbf{W} \leq 0 \ \forall \mathbf{W} \tag{9}$$

$$\Rightarrow \frac{d\tilde{\mathcal{L}}}{d\mathbf{W}}(\mathbf{W})^T \Delta\mathbf{W} \leq 0 \ \forall \mathbf{W}, \tag{10}$$

where the first implication follows from the fact that $\sigma(\cdot)$ is differentiable and increasing (it has strictly positive derivative), and the second implication follows from the chain rule. This implies that plasticity rules ($\Delta\mathbf{W}$) and trained neural circuits ($\mathbf{W}^*$) can at most partially constrain the space of viable objective functions the system could be minimizing.

B Why can't the brain do explicit gradient descent?

We have provided one surefire way to decrease an objective function by modifying the parameters of a neural network—'simply' take small steps in the direction of the gradient of the loss (Section 2.1). To appreciate the challenges faced by theories of normative plasticity, it's important to understand why a biological system *could not* do this: in this section we will provide a simplified argument as to why gradient descent within multilayer neural networks produces *nonlocal* parameter updates, thus failing our most critical desideratum for a normative plasticity theory (Section 2.2). More detailed arguments for multilayer neural networks can be found here (Lillicrap et al., 2020), and descriptions of why gradient descent becomes even more implausible for recurrent neural networks trained with either backpropagation through time

(Werbos, 1990) or real-time recurrent learning (Williams and Zipser, 1989) can be found here (Marschall et al., 2020).

The 'weight transport problem' is the most basic reason that gradient descent is implausible for neural networks. Suppose that we have a stimulus-dependent network response, $\mathbf{r}(\mathbf{W}^{in}) = f(\mathbf{W}^{in}\mathbf{s})$, where $\mathbf{r}$ is an $N \times 1$ vector, and $\mathbf{W}^{in}$ is an $N \times N^s$ weight matrix mapping stimuli $\mathbf{s}$ into responses after a pointwise nonlinearity $f(\cdot)$. This network response is decoded into a network output, $o(\mathbf{W}^{in}, \mathbf{s}) = \mathbf{W}^{out}\mathbf{r}(\mathbf{W}^{in})$, where $\mathbf{W}^{out}$ is a $1 \times N$ vector mapping network responses into a scalar output. Now suppose for simplicity that our loss for a single stimulus example is given by:

$$\mathcal{L} = \frac{1}{2}\left(\hat{o} - o(\mathbf{W}^{in}, \mathbf{s})\right)^2. \tag{11}$$

This objective is trying to bring the stimulus-dependent network response $o(\mathbf{W}^{in}, \mathbf{s})$ close to the target output $\hat{o}$, and is zero if and only if $o = \hat{o}$. A reasonable hypothesis would be that the gradient of this objective function with respect to a synaptic weight, $\mathbf{W}^{in}_{ij}$, will produce a parameter update that is local: we will see that this is not true. Taking the gradient, we have:

$$\begin{aligned}
\frac{d}{d\mathbf{W}^{in}_{ij}}\mathcal{L} &= \frac{1}{2}\frac{d}{d\mathbf{W}^{in}_{ij}}\left(\hat{o} - o(\mathbf{W}^{in}, \mathbf{s})\right)^2 && (12)\\
&= (\hat{o} - o)\,\frac{d}{d\mathbf{W}^{in}_{ij}} o(\mathbf{W}^{in}, \mathbf{s}) && (13)\\
&= (\hat{o} - o)\,\mathbf{W}^{out}_i \frac{d}{d\mathbf{W}^{in}_{ij}} f_i(\mathbf{W}^{in}\mathbf{s}) && (14)\\
&= (\hat{o} - o)\,\mathbf{W}^{out}_i f_i'(\mathbf{W}^{in}\mathbf{s})\mathbf{s}_j. && (15)
\end{aligned}$$

Breaking down this final update, we can see three terms: an error, $(\hat{o} - o)$, the neuron's *output weight* $\mathbf{W}^{out}_i$, and an approximately Hebbian term $f_i'(\mathbf{W}^{in}\mathbf{s})\mathbf{s}_j$, which requires only a combination of pre- and post-synaptic activity. One might be tempted to organize the plasticity rule into a error feedback signal received by the neuron, scaled by a neuron-specific synaptic weight $\mathbf{W}^{out}_i$, and then combined with Hebbian coactivity to produce a synaptic update (Fig. S1a). This would have the form of a three-factor plasticity rule (Frémaux and Gerstner, 2016), combining weighted feedback with pre- and post-synaptic activity. However, the weight transport problem is as follows: $\mathbf{W}^{out}_i$ provides the strength of a synapse in the *feedforward* pathway—how could it possibly come to be that a feedback learning pathway would have access to the *same* synaptic weight? The answer is that there is no evidence for such a system of weight sharing across feedforward and feedback pathways in the brain, though there are many hypotheses about how such a system could, in theory, be approximated by a normative plasticity algorithm. This problem becomes more pronounced in multilayer networks, where the error signal must be propagated through many interconnected connectivity layers.

It is also worth noting two key differentiability assumptions inherent to this approach. For one, we assume not only that the loss function $\mathcal{L}$ is differentiable, but that some 'error calculating' part of the brain does differentiate it. This requires knowledge of what the desired network output should be $\hat{o}$, which for many real-world tasks is not possible. Second, we assume that the network activation function $f(\cdot)$ is differentiable. Since neurons typically emit binary spikes, this differentiability assumption is not necessarily valid, though several modern methods have circumvented this problem by using either stochastic neuron models (Williams, 1992; Dayan and Hinton, 1996) or by using clever optimization tricks (Bellec et al., 2020). In subsequent sections, we will outline two canonical algorithms that employ clever tricks to circumvent the weight transport problem.

C REINFORCE

In this section, we will provide a mathematical tutorial on the REINFORCE learning algorithm (Williams, 1992), which is a mechanism for updating the parameters in a stochastic neural network for reinforcement learning objective functions. Its chief advantages are twofold: first, it only requires you to be able to evaluate an objective function (i.e. the reward received on any given trial), not the gradient of the objective function with respect to the parameters (Fig. S1b). This is very useful in situations in which the relationship between rewards and network outputs is not clear to an agent, as would be the case in many reinforcement learning scenarios. Second, under a broad range of biologically reasonable assumptions about a neural network architecture, the parameter updates produced by this algorithm are 'local,' meaning the information required for a parameter update would reasonably be available to a synapse in the brain. This algorithm produces updates that are within the class of 'reward-modulated Hebbian plasticity rules.' The chief disadvantage of this algorithm is its comparative data-inefficiency relative to backpropagation. In practice, far more data samples (or equivalently, much lower learning rates) will be required to produce the same improvements in performance compared to backpropagation (Werfel et al., 2003).

The REINFORCE algorithm and minor variations appear in different fields with different names. It is useful to keep track of these alternative names, because they all use roughly the same derivation, with some improvements or field-specific modifications. In machine learning, the algorithm is often referred to as *node perturbation* (Richards et al., 2019; Lillicrap et al., 2020; Werfel et al., 2003), because it involves correlating fluctuations in neuron (node) activity with reward signals. In computational neuroscience, it is sometimes called *3-factor* or *reward-modulated Hebbian* plasticity (Frémaux and Gerstner, 2016), though REINFORCE is only one of several algorithms referred to by these blanket terms. In reinforcement learning, REINFORCE is often treated as a member of the more general class of *policy gradient* (Sutton and Barto, 2018) methods, which can be used to train any parameterized stochastic agent through reinforcement. Policy gradient methods need not commit to a neural network architecture, and are consequently not always local. Lastly, very

similar methods are used for fitting variational Bayesian models, and are in these contexts referred to as either *black box variational inference* (Ranganath et al., 2014) or *neural variational inference* (Mnih and Gregor, 2014).

In what follows, we will provide a brief derivation of the REINFORCE learning algorithm for a 1-layer feedforward neural network. We will then discuss the many extensions of the algorithm as well as its strengths and limitations as a normative plasticity model.

C.1 Network model

Most neural networks used in machine learning are deterministic. However, neurons in biological systems fluctuate across trials and stimulus presentations, so modeling them as stochastic is often more appropriate. It will turn out that these fluctuations can be used to produce parameter updates in a way that a deterministic system could not.

First, we will assume that there are stimuli drawn from some stimulus distribution, $p(\mathbf{s})$, and we will define the neural network response to a given stimulus drawn from this distribution as:

$$\mathbf{r} = f(\mathbf{W}^{in}\mathbf{s}) + \sigma\boldsymbol{\eta}, \tag{16}$$

where the $\boldsymbol{\eta}$ is the source of random fluctuations which, for simplicity, is drawn from a standard normal distribution ($\mathcal{N}(0,1)$). In this equation, $\mathbf{s}$ is an $N_s \times 1$ vector, $\mathbf{W}^{in}$ is an $N_r \times N_s$ matrix, $f(\cdot)$ is the tanh nonlinearity, and $\boldsymbol{\eta}$ is an $N_r \times 1$ vector.

This equation defines a conditional probability distribution, $p(\mathbf{r}|\mathbf{s};\mathbf{W}^{in}) \sim \mathcal{N}(f(\mathbf{W}^{in}\mathbf{s}), \sigma^2)$. There is an interesting point here: neuron activities are now samples from this conditional probability distribution, and so we can study how neurons behave on average by taking expectations over the probability distribution.

For simplicity and clarity we will restrict ourselves to this neural architecture for our derivation, but the basic principles apply more generally to a variety of noise sources and neural architectures (see Section C.5).

C.2 Defining the objective

We will assume that our goal is to maximize some instantaneous reward $R(\mathbf{r},\mathbf{s})$ on average across many different samples of $R(\mathbf{r},\mathbf{s})$ and $\mathbf{s}$. This allows us to write our objective function $\mathcal{O}(\mathbf{W}^{in})$ as:

$$\mathcal{O}(\mathbf{W}^{in}) = \int R(\mathbf{r},\mathbf{s})p(\mathbf{r}|\mathbf{s};\mathbf{W}^{in})p(\mathbf{s})d\mathbf{r}d\mathbf{s}. \tag{17}$$

In practice, this integral might be analytically impossible to integrate, but we can always approximate it (because it is an expectation) using samples from $p(\mathbf{r}|\mathbf{s};\mathbf{W}^{in})$ and $p(\mathbf{s})$ as an empirical average over K samples $\mathbf{r}_k$ and $\mathbf{s}_k$:

$$\mathcal{O}(\mathbf{W}^{in}) \approx \frac{1}{K}\sum_{k=0}^{K} R(\mathbf{r}^{(k)},\mathbf{s}^{(k)}). \tag{18}$$

Procedurally, this would amount to sampling $\mathbf{s}$ and $\mathbf{r}$ each K times, calculating the reward for each trial, and taking an average.

C.3 Taking the gradient

Now that we have our objective function, we can evaluate its derivative with respect to a particular synapse $\mathbf{W}^{in}_{ij}$ in the network:

$$\frac{d\mathcal{O}(\mathbf{W}^{in})}{d\mathbf{W}^{in}_{ij}} = \frac{d}{d\mathbf{W}}\int R(\mathbf{r},\mathbf{s})p(\mathbf{r}|\mathbf{s};\mathbf{W}^{in})p(\mathbf{s})d\mathbf{r}d\mathbf{s} \tag{19}$$

$$= \int R(\mathbf{r},\mathbf{s})\left[\frac{d}{d\mathbf{W}^{in}_{ij}}p(\mathbf{r}|\mathbf{s};\mathbf{W}^{in})\right]p(\mathbf{s})d\mathbf{r}d\mathbf{s}. \tag{20}$$

We could theoretically stop here and evaluate $\frac{d}{d\mathbf{W}^{in}_{ij}}p(\mathbf{r}|\mathbf{s};\mathbf{W}^{in})$ explicitly. However, in the same way that we can approximate $\mathcal{O}(\mathbf{W}^{in})$ as an empirical average over samples, we would like to be able to approximate our derivative as an average. To do this requires us to keep our loss in the form of an expectation over $p(\mathbf{r}|\mathbf{s};\mathbf{W}^{in})p(\mathbf{s})$. We notice a convenient identity: $\frac{d}{d\mathbf{W}^{in}_{ij}}p(\mathbf{r}|\mathbf{s};\mathbf{W}^{in}) = \frac{d}{d\mathbf{W}^{in}_{ij}}\exp(\log p(\mathbf{r}|\mathbf{s};\mathbf{W}^{in})) = \left[\frac{d}{d\mathbf{W}^{in}_{ij}}\log p(\mathbf{r}|\mathbf{s};\mathbf{W}^{in})\right]p(\mathbf{r}|\mathbf{s};\mathbf{W}^{in})$, which is a simple application of the chain rule. Inserting this identity into the above equation, we get:

$$\frac{d\mathcal{O}(\mathbf{W}^{in})}{d\mathbf{W}^{in}_{ij}} = \int R(\mathbf{r},\mathbf{s})\left[\frac{d}{d\mathbf{W}^{in}_{ij}}\log p(\mathbf{r}|\mathbf{s};\mathbf{W}^{in})\right]p(\mathbf{r}|\mathbf{s};\mathbf{W}^{in})p(\mathbf{s})d\mathbf{r}d\mathbf{s} \tag{21}$$

$$\approx \frac{1}{K}\sum_{k=0}^{K} R(\mathbf{r}^{(k)},\mathbf{s}^{(k)})\left[\frac{d}{d\mathbf{W}^{in}_{ij}}\log p(\mathbf{r}^{(k)}|\mathbf{s}^{(k)};\mathbf{W}^{in})\right]. \tag{22}$$

Though this is an approximation, we note that by the Law of Large Numbers, we can improve its accuracy arbitrarily by increasing our number of samples K. In practice, however, taking $K = 1$ will prove to be the most straightforward way to

get an update that is local in time—although such an update will still on average match the true gradient exactly, its high variance can lead to very inefficient learning.

We have left the derivation completely general up until this point. Different choices of $p(\mathbf{r}|\mathbf{s};\mathbf{W})$ will produce different updates. Our particular choice gives:

$$\frac{d}{d\mathbf{W}^{in}_{ij}} \log p(\mathbf{r}|\mathbf{s};\mathbf{W}^{in}) = \frac{d}{d\mathbf{W}^{in}_{ij}} \sum_{i=0}^{N_r} \frac{1}{2\sigma^2}(\mathbf{r}_i - f_i(\mathbf{W}^{in}\mathbf{s}))^2 + C \tag{23}$$

$$= \frac{1}{\sigma^2} \sum_{n=0}^{N_r} (\mathbf{r}_i - f_i(\mathbf{W}^{in}\mathbf{s})) \frac{df_i(\mathbf{W}\mathbf{s})}{d\mathbf{W}^{in}_{ij}}. \tag{24}$$

For a particular weight $\mathbf{W}^{in}_{ij}$, $\frac{df_l(\mathbf{W}^{in}\mathbf{s})}{d\mathbf{W}_{ij}} = 0$ if $i \neq l$, so we have:

$$\frac{d}{d\mathbf{W}^{in}{}_{ij}} \log p(\mathbf{r}|\mathbf{s};\mathbf{W}) = \frac{1}{\sigma^2}(\mathbf{r}_i - f_i(\mathbf{W}\mathbf{s}))f_i'(\mathbf{W}\mathbf{s})\mathbf{s}_j. \tag{25}$$

Plugging this equation into Eq. 19 gives the following parameter update:

$$\Delta\mathbf{W}^{in}_{ij} \propto \frac{1}{K}\sum_{k=0}^{K} R(\mathbf{r}^{(k)},\mathbf{s}^{(k)}) \left[\frac{1}{\sigma^2}(\mathbf{r}^{(k)}_i - f_i(\mathbf{W}^{in}\mathbf{s}^{(k)}))f_i'(\mathbf{W}^{in}\mathbf{s}^{(k)})\mathbf{s}^{(k)}_j\right] \approx \frac{d\mathcal{O}(\mathbf{W}^{in})}{d\mathbf{W}^{in}_{ij}}. \tag{26}$$

If we want to update all of our parameters simultaneously using parallelized matrix operations, we can write this as an outer product:

$$\Delta\mathbf{W}^{in} \propto \frac{1}{K}\sum_{k=0}^{K} R(\mathbf{r}^{(k)},\mathbf{s}^{(k)}) \left[\frac{1}{\sigma^2}(\mathbf{r}^{(k)} - f(\mathbf{W}^{in}\mathbf{s}^{(k)})) \odot f'(\mathbf{W}^{in}\mathbf{s}^{(k)})\right] \mathbf{s}^{(k)T}, \tag{27}$$

where $\odot$ denotes a Hadamard (elementwise) vector product. Interestingly, the $\frac{1}{\sigma^2}(\mathbf{r} - f(\mathbf{W}^{in}\mathbf{s}))$ term here is exactly equal to $\boldsymbol{\eta}$.

C.4 Why don't we need the derivative of the loss?

One way of interpreting this parameter update is that neural units are correlating fluctuations in their neural activity with the rewards received to approximate $\frac{dR(\mathbf{r},\mathbf{s})}{d\mathbf{r}}$ (Fig. S1c). To see this, first notice that:

$$\mathbb{E}\left[b\left[\frac{1}{\sigma^2}(\mathbf{r}-f(\mathbf{W}^{in}\mathbf{s}))\odot f'(\mathbf{W}^{in}\mathbf{s})\right]\mathbf{s}^T\right]_{p(\mathbf{r}|\mathbf{s})}=0, \tag{28}$$

for any constant b, because $\mathbb{E}\left[\mathbf{r}-f(\mathbf{W}^{in}\mathbf{s})\right]_{p(\mathbf{r}|\mathbf{s})}=0$. If we take $b=\mathbb{E}\left[R(\mathbf{r},\mathbf{s})\right]_{p(\mathbf{r}|\mathbf{s})}$, then we can rewrite the gradient without changing its expected value:

$$\frac{d\mathcal{O}(\mathbf{W}^{in})}{d\mathbf{W}^{in}_{ij}}=\int\left(R(\mathbf{r},\mathbf{s})-\mathbb{E}\left[R(\mathbf{r},\mathbf{s})\right]_{p(\mathbf{r}|\mathbf{s})}\right)\left[\frac{1}{\sigma^2}(\mathbf{r}_i-f_i(\mathbf{W}^{in}\mathbf{s}))f_i'(\mathbf{W}^{in}\mathbf{s})\mathbf{s}_j\right]p(\mathbf{r}|\mathbf{s};\mathbf{W}^{in})p(\mathbf{s})d\mathbf{r}d\mathbf{s} \tag{29}$$

$$=\int\frac{1}{\sigma^2}Cov(R(\mathbf{r},\mathbf{s}),\mathbf{r}_i)\left[f_i'(\mathbf{W}^{in}\mathbf{s})\mathbf{s}_j\right]p(\mathbf{s})d\mathbf{s}, \tag{30}$$

where $Cov(R(\mathbf{r},\mathbf{s}),\mathbf{r}_i)=\int(R-\mathbb{E}\left[R\right]_{p(\mathbf{r}|\mathbf{s})})(\mathbf{r}_i-\mathbb{E}\left[\mathbf{r}_i\right]_{p(\mathbf{r}|\mathbf{s})})p(\mathbf{r}|\mathbf{s})d\mathbf{r}$ is the stimulus-conditioned covariance between network firing rates and reward. The sample-based parameter update is therefore using the fluctuations in neural activity to compute this covariance.

C.5 Assessing REINFORCE

Now that we have derived REINFORCE, we can examine its qualities as a normative plasticity theory. First, we ask: is this algorithm 'local' (Section 2.2)? The gradient for a particular synapse, $\frac{d\mathcal{O}(\mathbf{W}^{in})}{d\mathbf{W}^{in}_{ij}}$ can be approximated with samples in an environment with stimuli $\mathbf{s}$, firing rates $\mathbf{r}$, and rewards $R(\mathbf{r},\mathbf{s})$ by $R(\mathbf{r},\mathbf{s})\left[\frac{1}{\sigma^2}(\mathbf{r}_i-f_i(\mathbf{W}^{in}\mathbf{s}))f_i'(\mathbf{W}^{in}\mathbf{s})\mathbf{s}_j\right]$. To decide whether this could be a plasticity rule implemented (or more realistically, approximated) by a biological system, we need to think about what pieces of information a synapse would have to have available.

First, the synapse needs $\mathbf{s}_j$, which amounts to just the presynaptic input, a common feature of any Hebbian synaptic plasticity rule. Second, the synapse needs $\frac{1}{\sigma^2}(\mathbf{r}_i-f_i(\mathbf{W}^{in}\mathbf{s}))f_i'(\mathbf{W}^{in}\mathbf{s})$. $\frac{1}{\sigma^2}$ is a constant, and so can be absorbed into the learning rate. $\mathbf{r}_i$ is the postsynaptic firing rate, which is also a common feature of any Hebbian plasticity rule. $(\mathbf{W}^{in}\mathbf{s})_i$ is the current injected into the postsynaptic neuron, and $f_i(\cdot)$ and $f_i'(\cdot)$ are both monotonic functions of this current, so it is quite conceivable that these values could be approximated by a biochemical process. Third, every synapse needs access to the scalar reward value received on a given trial, $R(\mathbf{r},\mathbf{s})$. This is the most 'nonlocal' information involved in the parameter update, however, there exist many theories about how neuromodulatory systems in the brain can deliver information about reward diffusely to many synapses and induce plasticity (Section 2.2).

Now, we have already demonstrated that REINFORCE is able to perform approximate gradient descent for reinforcement learning objective functions—this in itself makes the algorithm very promising as a normative plasticity model (Section 2.1).

Its chief advantage is that it does not require detailed knowledge of the reward function $R(\mathbf{r}, \mathbf{s})$ (i.e. how to differentiate it), which means that an animal could simply receive a reward from its environment, and relay that reward signal diffusely to its synapses. However, this also restricts the types of objectives that could plausibly be learned by a neural system. Unsupervised learning objectives like the ELBO require detailed knowledge of every neural activity of every neuron in the circuit in order to be calculable (Appendix D), and there is no evidence for downstream neural circuits that perform such calculations. Therefore, even though in principle REINFORCE can be used to train a neural network on *any* objective, explicit reinforcement is much more plausible than other alternatives.

We have only provided a derivation for a single-layer rate-based neural network with additive Gaussian noise, but REINFORCE extends quite readily to multilayer (Williams, 1992), spiking (Frémaux et al., 2013), and recurrent networks (Miconi, 2017) without any loss of locality. This indicates that the algorithm is both architecture-general (Section 2.3) and can handle temporal environmental structure (Section 2.4). Further, because a weight update can be calculated in a single trial, animals could use it to learn online (Section 2.5). The biggest point of failure for REINFORCE is that it scales poorly with high complexity in stimuli or task, large numbers of neurons, or prolonged delays in receipt of reward (Werfel et al., 2003; Fiete, 2004; Bredenberg et al., 2021). The greater the number of neurons that contribute to reward and the higher the complexity of the reward function, the harder it becomes to estimate the correlation between a single neuron and reward, which is a prerequisite for the algorithm's function. Thus, though the algorithm is an unbiased estimator of the gradient, it can still be so variable an estimate as to be effectively useless in complex contexts. This suggests that if animals exploit the principles of REINFORCE to update synapses, it is likely an approach paired with other algorithms, or hybridized in a way that allows for better scalability.

The last way to assess REINFORCE is on the basis of how it can be tested (Section 2.7). The simplest way to test this algorithm is by examining whether scalar reward-like signals (i.e. $R(\mathbf{r}, \mathbf{s})$) have a multiplicative effect on local plasticity in a circuit. At a single-neuron level this corresponds to identifying neuromodulators that affect plasticity. At a feedback level this corresponds to identifying neuromodulatory systems that project to the circuit in question, and observing whether their stimulation or silencing improves or blocks circuit-level plasticity or behavioral learning performance, respectively. These steps do not identify REINFORCE as the only possibility, but it narrows down the field of possibilities considerably, removing all candidate algorithms that either do not require any feedback, or that require more detailed feedback signals (Fig. 3a).

D Wake-Sleep

Here we will provide a mathematical tutorial on the Wake-Sleep algorithm (Hinton et al., 1995; Dayan et al., 1995), which is one candidate biologically plausible learning

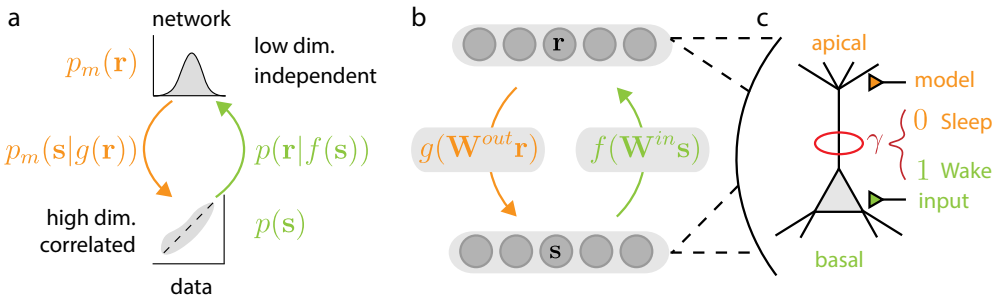

Figure S2: **The Wake-Sleep algorithm. a.** The four components of a good representation: $p_m(\mathbf{s}|\mathbf{r})$ and $p(\mathbf{r}|\mathbf{s})$ map $\mathbf{r}$ to $\mathbf{s}$ and back again from $\mathbf{s}$ to $\mathbf{r}$, respectively. $p_m(\mathbf{r})$ defines 'useful' features of a neural representation by constraining its topology. $p(\mathbf{s})$ provides the environmental input distribution, which the neural representation must match. **b.** The architecture of the Wake-Sleep algorithm: the decoder, $g(\mathbf{W}^{out}\mathbf{r})$ maps $\mathbf{r}$ to $\mathbf{s}$, and the forward map, $f(\mathbf{W}^{in}\mathbf{s})$ maps $\mathbf{s}$ to $\mathbf{r}$. **c.** Physically, these maps correspond to a multicompartmental pyramidal neuron model for each layer, where the 'model' synapses are on the apical dendrites, and the 'forward map' synapses are on the basal dendrites. γ gates which synapses determine neural activity, putting the network in the Wake phase $\gamma = 1$ or the Sleep phase $\gamma = 0$.

algorithm for constructing a representation in sensory cortices. We will first provide one possible formulation of representation learning as an optimization problem (Roweis and Ghahramani, 1999), and then introduce the Wake-Sleep algorithm[3], showing how the components necessary to the algorithm could be mapped onto a multicompartmental dendritic neuron model with local synaptic learning. We will then discuss how the algorithm can be extended beyond our simplified introduction.

D.1 Defining a good objective

Suppose that at any given moment in time, a neural network is receiving sensory stimuli $\mathbf{s}$ from its environment. Our first challenge is to articulate what it would mean to form a good neural representation $\mathbf{r}$ of these stimuli (Fig. S2a). First of all, 'represented' stimuli should be decodable from neural firing rates, i.e. there should exist a mapping $g(\cdot) : \mathbb{R}^{N^r} \to \mathbb{R}^{N^s}$ such that $\mathbf{s} \approx g(\mathbf{r})$. Second, we will also argue that neural firing rates should be decodable from *stimuli*, i.e. there should exist a mapping $f(\cdot) : \mathbb{R}^{N^s} \to \mathbb{R}^{N^r}$ such that $\mathbf{r} \approx f(\mathbf{s})$—this means that there cannot be 'extra' features of neural activity that are not contained within the stimuli themselves. This amounts to postulating an approximately bijective relationship between stimuli and firing rates. It means that neural activities should directly correspond to stimuli that have been received.

If these two requirements were sufficient, we might want to simply have one neuron per stimulus dimension, and have it faithfully replicate its immediate input as accu-

[3]For another excellent tutorial with more of a machine learning focus, see (Kirby, 2006).

rately as possible, i.e. we would take $f(\mathbf{s}) = \mathbb{I}\mathbf{s}$ and $g(\mathbf{r}) = \mathbb{I}\mathbf{r}$, where $\mathbb{I}$ is an identity matrix, so that $\mathbf{r} = \mathbb{I}\mathbf{s} = \mathbf{s}$. This identity transformation is obviously not useful, which makes one wonder—what does it mean for a transformation to be useful? Most, if not all unsupervised machine learning and neuroscientific conceptions of a 'useful' representation reduce to some formulation of either metabolic or coding efficiency. Approaches within this 'efficiency' umbrella include dimensionality reduction (Roweis and Ghahramani, 1999), clustering (Illing et al., 2021; Dayan et al., 1995), gain control (Simoncelli and Heeger, 1998), whitening/factorization (Rezende et al., 2014), and sparsity (Simoncelli and Olshausen, 2001). Each of these definitions of 'usefulness' can be formulated as statements about the distribution of neural activities, independent of particular received stimuli, e.g. there are fewer neurons than stimulus dimensions (dimensionality reduction), neural activations occupy roughly discrete clusters in state space (clustering), neurons tend to be uncorrelated with one another (whitening/factorization), or neurons typicaly have low, sparse firing rates (gain control/sparsity/metabolic efficiency). In our formulation, ultimately learning will be unsupervised because we have made *a priori* determinations of what constitutes an efficient representation, and seek to transform incoming data to match those determinations.

Under our definition outlined so far, there are four components of a representation: the stimuli $\mathbf{s}$ themselves, distributed according to some probability distribution $p(\mathbf{s})$ determined by the environment; a decoder, which we will formulate probabilistically as $p_m(\mathbf{s}|g(\mathbf{r};\theta_m))$, which models the probability of $\mathbf{s}$ given our mapping from neural firing rates $\mathbf{r}$; a forward mapping from $\mathbf{s}$ to $\mathbf{r}$, which we will also formulate probabilistically as $p(\mathbf{r}|f(\mathbf{s};\theta))$; and our definition of efficiency, which dictates how neural firing rates 'should' be distributed, independently of stimuli themselves $p_m(\mathbf{r})$. Notice that here we have parameterized the forward map $p(\mathbf{r}|f(\mathbf{s};\theta))$ and the decoder (inverse map) $p_m(\mathbf{s}|g(\mathbf{r};\theta_m))$: once we formulate our objective, these will be the parameters that are adjusted to minimize it. $p(\mathbf{s})$—the environmental data distribution—obviously cannot change, but we could (and in practice would often want to) parameterize $p_m(\mathbf{r})$ and also fit those parameters. We have formulated our four components using probability distributions: after describing our objective function in these terms, we will show one possible way of mapping the components onto neural architecture.

Now, we have evocatively organized our components into two groups: $p_m(\mathbf{s}|g(\mathbf{r};\theta_m))$ and $p_m(\mathbf{r})$, versus $p(\mathbf{s})$ and $p(\mathbf{r}|f(\mathbf{s};\theta))$. The first group forms a joint distribution $p_m(\mathbf{r},\mathbf{s};\theta_m)$ which has the subscript m to indicate that it is a generative *model* of the data. Ideally, if its parameters were accurately fit, we could sample $\mathbf{r} \sim p_m(\mathbf{r})$, and then sample $\mathbf{s} \sim p_m(\mathbf{s}|g(\mathbf{r};\theta_m))$ and get a stimulus that looks like realistic environmental data. The second group also forms a joint distribution $p(\mathbf{r},\mathbf{s};\theta)$, which amounts to a forward mapping: we could receive a stimulus from the environment, and then have the probability distribution for firing rates $\mathbf{r}$ that correspond to it. Organizing our models in this way will allow us to achieve biophysical realism: $g(\cdot;\theta)$ and $f(\cdot;\theta)$ will correspond to actual synaptic connections in a model neural network. In practice, ordinary perception as we traditionally conceive of it would correspond to the forward mapping $f(\cdot;\theta)$. Interestingly, at the end of our derivation, it will

become clear how an additional representational feature, 'detachability' (Clark and Toribio, 1994)—a mechanism to activate neurons in the absence of the sensory stimuli that correspond to them—will be an emergent property of our formulation. We will show how a neural system might be able to leverage the $g(\cdot;\theta_m)$ to accomplish 'detachment', which one might imagine mapping perceptually to imagination, planning, prediction, hallucination, or possibly dreaming in different contexts.

For our representation to be good, the forward map should match its inverse, i.e. $p(\mathbf{r},\mathbf{s};\theta) \approx p_m(\mathbf{r},\mathbf{s};\theta_m)$. We could imagine formulating many objective functions that could accomplish this goal, but most of them will not accommodate an approximate optimization algorithm that will end up corresponding to a viable normative plasticity model. We will select the Kullback-Liebler (KL) divergence between these two distributions, precisely because it will produce such a normative plasticity model. Notice, though our presentation of the derivation is top-down, it is disingenuous to characterize normative plasticity model development strictly as top-down: locality would not magically emerge from an arbitrary choice of objective function, but rather this choice of objective function is superior to its many alternatives only *because* it produces locality (we won't be able to see why locality emerges until after we have defined p and p_m explicitly and have derived parameter updates). We take our objective function to be:

$$\begin{aligned}\mathcal{L}_{Wake} &= D_{KL}(p(\mathbf{r},\mathbf{s};\theta)||p_m(\mathbf{r},\mathbf{s};\theta_m)) \\ &= \int \ln\left(\frac{p(\mathbf{r},\mathbf{s};\theta)}{p_m(\mathbf{r},\mathbf{s};\theta_m)}\right) p(\mathbf{r},\mathbf{s};\theta) d\mathbf{r} d\mathbf{s}. \end{aligned} \tag{31}$$

We have evocatively named this loss $\mathcal{L}_{Wake}$ because we will be optimizing this objective function during the Wake phase of the algorithm. We will also later appeal to the opposite KL divergence, which we will be optimizing during the Sleep phase:

$$\begin{aligned}\mathcal{L}_{Sleep} &= D_{KL}(p_m(\mathbf{r},\mathbf{s};\theta_m)||p(\mathbf{r},\mathbf{s};\theta)) \\ &= \int \ln\left(\frac{p_m(\mathbf{r},\mathbf{s};\theta_m)}{p(\mathbf{r},\mathbf{s};\theta)}\right) p_m(\mathbf{r},\mathbf{s};\theta_m) d\mathbf{r} d\mathbf{s}. \end{aligned} \tag{32}$$

These objectives share a global minimum ($p_m = p$), if it exists, but are not the same objective function, because unlike a traditional distance metric, the KL divergence is not symmetric. However, *near* the global minimum, they become approximately equivalent (Dayan et al., 1995; Bredenberg et al., 2021), which will be an important consideration in assessing the convergence properties of the Wake-Sleep algorithm. Unlike REINFORCE, which will work for any reward function $R(\mathbf{r},\mathbf{s})$, the Wake-Sleep algorithm will only work for objectives formulated in this way: in this case the choice of objective function is intimately related to the resultant plasticity rule.

D.1.1 Equivalence to the Evidence Lower Bound*

It should be noted that $\mathcal{L}_{Wake}$ has a long history in unsupervised machine learning, and does not always appear in the context of training a sensory representational system through normative plasticity. In fact, minimizing $\mathcal{L}_{Wake}$ is equivalent to minimizing the variational free energy or maximizing the evidence lower bound (ELBO), the objective underlying the variational autoencoder (Rezende et al., 2014; Kingma and Welling, 2014) and the Expectation-Maximization algorithm for latent state models (Roweis and Ghahramani, 1999). Here, to help relate to the broader literature, we will elaborate on this equivalence for the interested reader. This section is a technical aside, which the uninterested reader may safely skip. In traditional machine learning terms, as we will see, the $\mathcal{L}_{Wake}$ objective is equivalent to maximizing the ELBO, and will fit a generative model $p_m(\mathbf{r}, \mathbf{s}; \theta_m)$ to data, as well as train a forward map $p(\mathbf{r}|\mathbf{s}; \theta)$ to perform approximate Bayesian inference with respect to that model (i.e. we want $p(\mathbf{r}|\mathbf{s}; \theta) \approx p_m(\mathbf{r}|\mathbf{s}; \theta_m)$).

To fit a generative model to data, we would typically use maximum likelihood estimation: we would find the parameters of our generative model $p_m(\mathbf{r}, \mathbf{s}; \theta_m)$ that match the distribution of data points as accurately as possible by minimizing with respect to θ:

$$D_{KL}(p(\mathbf{s})||p_m(\mathbf{s}; \theta_m)) = \int \ln\left(\frac{p(\mathbf{s})}{p_m(\mathbf{s}; \theta_m)}\right) p(\mathbf{s}) d\mathbf{s}. \tag{33}$$

When this objective is 0, samples drawn from $p_m(\mathbf{s}; \theta_m)$ will be indistinguishable from samples drawn from $p(\mathbf{s})$, indicating that we have an accurate model of the data distribution. But we are not only interested in fitting a generative model: when our network receives a stimulus $\mathbf{s}$, we would like it to infer the probability distribution over latent representational states that could correspond to that stimulus, $p_m(\mathbf{r}|\mathbf{s}; \theta_m)$. However, we haven't defined this quantity, only $p_m(\mathbf{r})$ and $p_m(\mathbf{s}|\mathbf{r}; \theta_m)$. From a purely machine learning perspective, we might just try to compute $p_m(\mathbf{r}|\mathbf{s}; \theta_m)$ explicitly using Bayes' Theorem:

$$p_m(\mathbf{r}|\mathbf{s}; \theta_m) = \frac{p_m(\mathbf{r}) p_m(\mathbf{s}|\mathbf{r}; \theta)}{\int p_m(\mathbf{r}) p_m(\mathbf{s}|\mathbf{r}; \theta) d\mathbf{r} d\mathbf{s}}, \tag{34}$$

and for simple generative models this might work. However, for complex, nonlinear models, calculating the high-dimensional integral in the denominator analytically is impossible, and approximating it through Monte Carlo methods is time consuming to the point of intractability. This is motivation enough for machine learning applications, but further, it is not clear how biological system could compute such an integral rapidly upon receiving a single stimulus. So instead, we might try a different approach. We can take our explicitly defined and parameterized forward map $p(\mathbf{r}|\mathbf{s}; \theta)$ and train it to approximate $p_m(\mathbf{r}|\mathbf{s}; \theta_m)$ as closely as possible by minimizing the expected KL divergence:

$$\mathbb{E}\left[D_{KL}(p(\mathbf{r}|\mathbf{s};\theta)||p_m(\mathbf{r}|\mathbf{s};\theta_m))\right]_{p(\mathbf{s})} = \int \ln\left(\frac{p(\mathbf{r}|\mathbf{s};\theta)}{p_m(\mathbf{r}|\mathbf{s};\theta_m)}\right) p(\mathbf{r}|\mathbf{s};\theta)p(\mathbf{s})d\mathbf{r}d\mathbf{s}. \quad (35)$$

If objective is approximately 0, then we do not need to perform Bayes' theorem to calculate the posterior $p_m(\mathbf{r}|\mathbf{s};\theta_m)$, because we have access to a perfect (or near-perfect) approximation $p(\mathbf{r}|\mathbf{s};\theta)$ that we can calculate explicitly or sample from. If $p(\mathbf{r}|\mathbf{s};\theta)$ is parameterized appropriately, this is usually much easier, and potentially could be implemented by a neural network. Now we have two objectives that we want to minimize: one to fit our generative model, and the other to perform approximate inference. It seems natural to add them and minimize them jointly. First, we notice that adding our second objective defines the following inequality:

$$D_{KL}(p(\mathbf{s})||p_m(\mathbf{s};\theta_m)) \leq D_{KL}(p(\mathbf{s})||p_m(\mathbf{s};\theta_m)) + \mathbb{E}\left[D_{KL}(p(\mathbf{r}|\mathbf{s};\theta)||p_m(\mathbf{r}|\mathbf{s};\theta_m))\right]_{p(\mathbf{s})}. \quad (36)$$

due to the positivity of the KL divergence. Second, we note that adding these two objectives together really just gives us $\mathcal{L}_{Wake}$:

$$D_{KL}(p(\mathbf{s})||p_m(\mathbf{s};\theta_m)) + \mathbb{E}\left[D_{KL}(p(\mathbf{r}|\mathbf{s};\theta)||p_m(\mathbf{r}|\mathbf{s};\theta_m))\right]_{p(\mathbf{s})} = D_{KL}(p(\mathbf{r},\mathbf{s};\theta)||p_m(\mathbf{r},\mathbf{s};\theta)) \quad (37)$$

$$= \mathcal{L}_{Wake}, \quad (38)$$

where the first equality follows from adding Eqs. 33 and 35 and using the properties of the logarithm and expectations.

This alternative construction demonstrates that minimizing that our objective function $\mathcal{L}_{Wake}$ trains our system to perform two separate model-fitting functions: training a generative model and training an approximate inference distribution. From here we can also see its equivalence to the variational free energy and the ELBO:

$$D_{KL}(p(\mathbf{s})||p_m(\mathbf{s};\theta_m)) \leq \mathcal{L}_{Wake} \quad (39)$$

$$= \int \ln\left(\frac{p(\mathbf{r},\mathbf{s};\theta)}{p_m(\mathbf{r},\mathbf{s};\theta_m)}\right) p(\mathbf{r},\mathbf{s};\theta)d\mathbf{r}d\mathbf{s} \quad (40)$$

$$= \int \ln\left(\frac{p(\mathbf{r}|\mathbf{s};\theta)}{p_m(\mathbf{r},\mathbf{s};\theta_m)}\right) p(\mathbf{r},\mathbf{s};\theta)d\mathbf{r}d\mathbf{s} + \int \left(\ln p(\mathbf{s})\right) p(\mathbf{r}|\mathbf{s};\theta)p(\mathbf{s})d\mathbf{r}d\mathbf{s} \quad (41)$$

$$= \int \ln\left(\frac{p(\mathbf{r}|\mathbf{s};\theta)}{p_m(\mathbf{r},\mathbf{s};\theta_m)}\right) p(\mathbf{r},\mathbf{s};\theta)d\mathbf{r}d\mathbf{s} + \int \left(\ln p(\mathbf{s})\right) p(\mathbf{s})d\mathbf{s}. \quad (42)$$

Now, by definition $D_{KL}(p(\mathbf{s})||p_m(\mathbf{s};\theta_m)) = \int \left(\ln p(\mathbf{s})\right) p(\mathbf{s})d\mathbf{s} - \int \left(\ln p_m(\mathbf{s};\theta_m)\right) p(\mathbf{s})d\mathbf{s}$, the first term of which also appears on the right hand side of our inequality. Furthermore, $\int \left(\ln p(\mathbf{s})\right) p(\mathbf{s})d\mathbf{s}$ is not a function θ_m or θ, so from the perspective of

optimization, it is an irrelevant additive constant. We subtract it from both sides to get:

$$-\int (\ln p_m(\mathbf{s};\theta_m))\, p(\mathbf{s}) d\mathbf{s} \leq \int \ln\left(\frac{p(\mathbf{r}|\mathbf{s};\theta)}{p_m(\mathbf{r},\mathbf{s};\theta_m)}\right) p(\mathbf{r},\mathbf{s};\theta) d\mathbf{r} d\mathbf{s}. \tag{43}$$

This expression on the left is the negative log-likelihood, and the expression on the right is the variational free energy, which is the negative of the ELBO. This shows that $\mathcal{L}_{Wake}$ and the variational free energy differ only by an additive constant from the perspective of optimization: minimizing one is the same as minimizing the other. Similarly, $\mathcal{L}_{Sleep}$ corresponds to an upper bound on the reverse KL divergence, $D_{KL}(p_m(\mathbf{s};\theta_m)||p(\mathbf{s}))$.

D.2 Defining p and p_m

Let us start by selecting three features of our representation that we think will be useful, i.e. efficient. First, we want our neurons to be metabolically efficient: a biological system cannot have neurons wasting energetic resources by firing too much (Simoncelli, 2003). One way of requiring this would be to stipulate that the squared norm of our neural firing rate vector, $\|\mathbf{r}\|_2^2$ lies within some reasonable range of activation values. Second, we want to reduce the dimensionality of our representation: many naturalistic datasets are low-dimensional, and it may be wasteful to represent some high-dimensional features of stimuli that are just due to sensor noise. To accomplish this, we will stipulate that $N_r \ll N_s$, where N_r is the representation's dimensionality, and N_s is the stimulus dimension. Third, we will require that individual neural activations should be independent from one another, which will allow individual neurons to extract important features of the data without requiring full knowledge of the activity of other neurons in the representation. To achieve a representation that embodies these three desired features, we define $p_m(\mathbf{r})$ as follows:

$$p_m(\mathbf{r}) \sim \mathcal{N}(0,1), \tag{44}$$

i.e. we will require that the representation, averaged over stimuli, will match an N_r-dimensional multivariate normal distribution, where individual axes $\mathbf{r}_i$ are independent from one another (uncorrelated), and where the normal distribution naturally restricts the probable range of neural activities to lie within bounds determined by the variance (arbitrarily set to 1). Though this distribution captures several intuitions for how neural representations should function, it is clearly a toy model for several reasons: it does not restrict firing rates to be positive, it does not allow for activities to be discrete spikes, it does not account for temporal dynamics, etc. We will discuss later how each of these extensions have been done before, but for now, many features of our model $p_m(\mathbf{r},\mathbf{s})$ and our forward map $p(\mathbf{r}|\mathbf{s})$ will be unrealistic for didactic purposes.

Now we define the probabilistic decoder $p_m(\mathbf{s}|\mathbf{r})$ (Fig. S2b), which takes neural firing rates and produces estimates of stimuli, as follows:

$$p_m(\mathbf{s}|\mathbf{r};\mathbf{W}^{out}) \sim \mathcal{N}(g(\mathbf{W}^{out}\mathbf{r}),\sigma_s), \tag{45}$$

where $g(\cdot)$ is an arbitrary nonlinearity, and σ_s^2 is the variance of the decoder. In this probability distribution, we will treat the $N_s \times N_r$ matrix $\mathbf{W}^{out}$ as a free parameter which we will train to optimize our objective.

Similarly, we can define the forward map $p(\mathbf{r}|\mathbf{s})$, which takes environmental stimuli and produces firing rates, as follows:

$$p(\mathbf{r}|\mathbf{s}; \mathbf{W}^{in}) = \mathcal{N}(f(\mathbf{W}^{in}\mathbf{s}), \sigma_r), \tag{46}$$

where $f(\cdot)$ is an arbitrary (potentially different) nonlinearity, and σ_r will ultimately correspond to intrinsic neural variability. Here, the $N_r \times N_s$ matrix $\mathbf{W}^{in}$ is the free parameter. Thus, $\mathbf{W}^{in}$ and $\mathbf{W}^{out}$, are the free parameters in our simple construction.

We have not yet made clear how these parameters and functions could map onto an actual neural architecture: we will do this after defining the learning algorithm, so that it is clear what the necessary components of the algorithm are. Interestingly, we do not have to define $p(\mathbf{s})$ at all. This distribution is determined by the environment. In fact, a learning system should ideally be as agnostic as possible to the specific form of $p(\mathbf{s})$ as possible, in order to be able to adapt strange and unforeseen changes in the statistics of the world. The Wake-Sleep algorithm is ideal in that it makes little-to-no assumption about $p(\mathbf{s})$, but as we will see, it may perform poorly if it is not possible to obtain a close match between p and p_m. This might occur if the environmental distribution of $\mathbf{s}$ is much higher dimensional than the number of neurons, or is in some other way more complex than the generative model.

D.3 Approximating the loss gradient

Having defined our objective function and probability distributions p and p_m, we can now derive the Wake-Sleep algorithm. First, we will show that we can obtain a promising update for $\mathbf{W}^{out}$ by performing gradient descent on $\mathcal{L}_{Wake}$ (the Wake phase of learning). We will next show that we can obtain a similarly promising update for $\mathbf{W}^{in}$ by performing gradient descent on $\mathcal{L}_{Sleep}$ (the Sleep phase of learning). One might easily wonder why we did not perform gradient descent on $\mathcal{L}_{Wake}$ with respect to $\mathbf{W}^{in}$, instead of $\mathcal{L}_{Sleep}$: we will next show why it would be a bad idea to do this. Lastly, we will describe two perspectives on how these resultant updates can be viewed as a unified form of approximate optimization.

D.3.1 Wake

We start by calculating the negative gradient of $\mathcal{L}_{Wake}$ with respect to a particular parameter $\mathbf{W}_{ij}^{out}$ from $p_m(\mathbf{s}|\mathbf{r}; \mathbf{W}^{out})$:

$$-\frac{d\mathcal{L}_{Wake}}{d\mathbf{W}_{ij}^{out}} = -\frac{d}{d\mathbf{W}_{ij}^{out}} \int \ln\left(\frac{p(\mathbf{r},\mathbf{s};\mathbf{W}^{in})}{p_m(\mathbf{r},\mathbf{s};\mathbf{W}^{out})}\right) p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) d\mathbf{r}d\mathbf{s} \tag{47}$$

$$= -\frac{d}{d\mathbf{W}_{ij}^{out}} \int \left[\ln p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) - \ln p_m(\mathbf{s}|\mathbf{r};\mathbf{W}^{out}) - \ln p_m(\mathbf{r})\right] p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) d\mathbf{r}d\mathbf{s} \tag{48}$$

$$= -\int \frac{d}{d\mathbf{W}_{ij}^{out}} \left[\ln p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) - \ln p_m(\mathbf{s}|\mathbf{r};\mathbf{W}^{out}) - \ln p_m(\mathbf{r})\right] p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) d\mathbf{r}d\mathbf{s} \tag{49}$$

$$= \int \left[\frac{d}{d\mathbf{W}_{ij}^{out}} \ln p_m(\mathbf{s}|\mathbf{r};\mathbf{W}^{out})\right] p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) d\mathbf{r}d\mathbf{s} \tag{50}$$

Plugging in the probability density function for $p_m(\mathbf{s}|\mathbf{r};\mathbf{W}^{out}) \sim \mathcal{N}(g(\mathbf{W}^{out}\mathbf{r}), \sigma_s^2)$, we end up with:

$$-\frac{d\mathcal{L}_{Wake}}{d\mathbf{W}_{ij}^{out}} = \int \left[\frac{d}{d\mathbf{W}_{ij}^{out}} \frac{1}{2\sigma_s^2} \sum_{i=0}^{N_s} (\mathbf{s} - g(\mathbf{W}^{out}\mathbf{r}))^2\right] p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) d\mathbf{r}d\mathbf{s}. \tag{51}$$

Similar to our derivation for REINFORCE, we see that for a particular weight $\mathbf{W}_{ij}^{out}$, $\frac{dg_l(\mathbf{W}^{out}\mathbf{r})}{d\mathbf{W}_{ij}^{out}} = 0$ if $i \neq l$. Thus, we have:

$$-\frac{d\mathcal{L}_{Wake}}{d\mathbf{W}_{ij}^{out}} = \int \frac{1}{\sigma_s^2} \left[(\mathbf{s}_i - g_i(\mathbf{W}^{out}\mathbf{r}))g_i'(\mathbf{W}^{out}\mathbf{r})\mathbf{r}_j\right] p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) d\mathbf{r}d\mathbf{s}. \tag{52}$$

Again, similar to REINFORCE, we can approximate this update as the network actively 'perceives': we receive a sampled environmental stimulus $\mathbf{s}^{(k)}$, and then sample from the probability distribution $p(\mathbf{r}|\mathbf{s}^{(k)};\mathbf{W}^{in})$ to obtain a firing rate sample $\mathbf{r}^{(k)}$. Then across K samples, we calculate the approximate parameter update:

$$\Delta\mathbf{W}_{ij}^{out} \propto \frac{1}{\sigma_s^2 K} \sum_{k=0}^{K} \left[(\mathbf{s}_i^{(k)} - g_i(\mathbf{W}^{out}\mathbf{r}^{(k)}))g_i'(\mathbf{W}^{out}\mathbf{r}^{(k)})\mathbf{r}_j^{(k)}\right] \approx -\frac{d\mathcal{L}_{Wake}}{d\mathbf{W}_{ij}^{out}}. \tag{53}$$

If we want learning to be able to occur online (Section 2.5), then we can take $K = 1$, and sacrifice some precision of our estimate. This update has the form of a prediction error, where the the error between the true stimulus $\mathbf{s}_i^{(k)}$ and the network's decoded estimate $g_i(\mathbf{W}^{out}\mathbf{r}^{(k)})$ combine with presynaptic inputs $r_j^{(k)}$ to produce parameter updates. In Section D.4 we will analyze in detail how this parameter update could correspond to a local synaptic update for a particular neuron model.

D.3.2 Sleep

So far, other than performing stochastic gradient descent over K samples, we have introduced no approximation into our algorithm. We might be tempted to perform gradient descent on $\mathcal{L}_{Wake}$ with respect to $\mathbf{W}^{in}$ too: though we will defer the discussion of this point for later, it turns out to be a bad idea (see Section D.4.1). Instead, we will perform an *almost identical* procedure, but perform gradient descent on $\mathcal{L}_{Sleep}$ instead. As discussed in Section 2.1, one way of interpreting this change in loss is that we now have two different sets of parameters (i.e. synapses) in our system, $\mathbf{W}^{in}$ and $\mathbf{W}^{out}$ which are optimizing two different, albeit closely related objectives, $\mathcal{L}_{Sleep}$ and $\mathcal{L}_{Wake}$, respectively. An alternative perspective that we will discuss is that $\mathbf{W}^{in}$ is also optimizing $\mathcal{L}_{Wake}$, but is only performing an approximate gradient descent. We will discuss in Section D.4.2 how this added complexity affects the convergence and quality of the algorithm. Starting with $\mathcal{L}_{Sleep}$, we have:

$$-\frac{d\mathcal{L}_{Sleep}}{d\mathbf{W}^{in}_{ij}} = -\frac{d}{d\mathbf{W}^{in}_{ij}} \int \ln \left(\frac{p_m(\mathbf{r}, \mathbf{s}; \mathbf{W}^{out})}{p(\mathbf{r}, \mathbf{s}; \mathbf{W}^{in})} \right) p_m(\mathbf{r}, \mathbf{s}; \mathbf{W}^{out}) d\mathbf{r} d\mathbf{s} \tag{54}$$

$$= \int \left[\frac{d}{d\mathbf{W}^{in}_{ij}} \frac{1}{2\sigma_r^2} \sum_{i=0}^{N_r} (\mathbf{r} - f(\mathbf{W}^{in}\mathbf{s}))^2 \right] p_m(\mathbf{r}, \mathbf{s}; \mathbf{W}^{out}) d\mathbf{r} d\mathbf{s}, \tag{55}$$

where we have followed exactly the same steps as in Eqs. 47-51.

As before, we notice that for a particular weight $\mathbf{W}^{in}_{ij}$, $\frac{df_l(\mathbf{W}^{in}\mathbf{s})}{d\mathbf{W}^{in}_{ij}} = 0$ if $i \neq l$. Thus, we have:

$$-\frac{d\mathcal{L}_{Sleep}}{d\mathbf{W}^{in}_{ij}} = \int \frac{1}{\sigma_r^2} \left[(\mathbf{r}_i - f_i(\mathbf{W}^{in}\mathbf{s})) f_i'(\mathbf{W}^{in}\mathbf{s}) \mathbf{s}_j \right] p_m(\mathbf{r}, \mathbf{s}; \mathbf{W}^{out}) d\mathbf{r} d\mathbf{s}. \tag{56}$$

Now we can approximate this update with samples from $p_m(\mathbf{r}, \mathbf{s}; \mathbf{W}^{out})$. Notice that we are no longer actively perceiving via the forward mapping $p(\mathbf{r}|\mathbf{s})$ in response to sampled environmental stimuli. Instead, activity is first internally generated via $\mathbf{r}^{(k)} \sim p_m(\mathbf{r})$, before propagating to the stimulus layer to produce artificial stimuli via $\mathbf{s}^{(k)} \sim p_m(\mathbf{s}|\mathbf{r}^{(k)}; \mathbf{W}^{out})$. This is termed the Sleep phase of the algorithm evocatively: an animal could not perform this type of learning while actively moving through an environment, and if it did perceive, such percepts would appear hallucinatory or dream-like, being reflective of the animal's model rather than reality. Given our K samples, we calculate the approximate parameter update:

$$\Delta\mathbf{W}^{in}_{ij} \propto \frac{1}{\sigma_r^2 K} \sum_{k=0}^{K} \left[(\mathbf{r}_i^{(k)} - f_i(\mathbf{W}^{in}\mathbf{s}^{(k)})) f_i'(\mathbf{W}^{in}\mathbf{s}^{(k)}) \mathbf{s}_j^{(k)} \right] \approx -\frac{d\mathcal{L}_{Sleep}}{d\mathbf{W}^{in}_{ij}}. \tag{57}$$

Now, this update should look almost equivalent to the Wake update for $\mathbf{W}^{out}$ (Eq. 53). As with the Wake update, if we want learning to occur online we can take $K = 1$. It turns out that the variability induced by this sampled approximation is *much* less than the variability induced by the REINFORCE algorithm, and is the chief reason for its superior performance and scalability (Bredenberg et al., 2021). However, it is very important to note that we are sampling from p_m instead of p. Because our two parameter updates, Eq. 53 and Eq. 57 require sampling from two different probability distributions and individual neurons $\mathbf{r}$ could only be sampling from one probability distribution at a time, the updates are necessarily computed during different *phases*. The Wake-Sleep algorithm consists of alternating between sampling from p to compute updates for $\mathbf{W}^{out}$ (the Wake phase; Eq. 53) and sampling from p_m to compute updates for $\mathbf{W}^{in}$ (the Sleep phase; Eq. 57). As we discuss in Section D.4, we should be appropriately cautious about what these alternative phases could possibly mean for a biological organism.

D.4 Assessing Wake Sleep

Having derived our Wake-Sleep parameter updates, we are finally in a position to assess the degree to which it satisfies our desiderata. We have provided a very simplified derivation of the Wake-Sleep algorithm, for a single-layer rate-based network. However, the algorithm generalizes well to recurrent, spiking, and multilayer architectures (Dayan and Hinton, 1996) (Section 2.3), and these modifications do make the algorithm more realistic as a normative plasticity model. However, it will still be very useful to show how the various components of the algorithm as we have derived it could potentially map onto realistic biological structures (Fig. S2c). First of all, we observe that both $\mathbf{s}$ and $\mathbf{r}$ need to be able to sample from either p_m or p—for this to be possible, $\mathbf{s}$ must be *internal* to the brain, since sampling from p_m affects both $\mathbf{r}$ and $\mathbf{s}$ simultaneously and would have to occur while an animal is not consciously acting in its environment. Therefore, it is best to think of $\mathbf{s}$ as a stimulus layer of neurons, and of $\mathbf{r}$ as a downstream layer of neurons receiving feedforward inputs. Next, we suppose that there is a global gating signal γ that determines the phase of the network— if $\gamma = 1$, the network is in the Wake phase, and if $\gamma = 0$, the network is in the Sleep phase. Now we observe that the following equations will produce valid samples:

$$\mathbf{r} = \gamma f(\mathbf{W}^{in}\mathbf{s}) + (\gamma\sigma_r + (1-\gamma))\boldsymbol{\eta}_r \tag{58}$$

$$\mathbf{s} = \gamma\mathbf{s}_p + (1-\gamma)(g(\mathbf{W}^{out}\mathbf{r}) + \sigma_s\boldsymbol{\eta}_s), \tag{59}$$

where $\mathbf{s}_p \sim p(\mathbf{s})$ is an incoming sensory input, and $\boldsymbol{\eta}_s, \boldsymbol{\eta}_r \sim \mathcal{N}(0,1)$ are sources of intrinsic noise for neurons in the stimulus, and downstream layers, respectively. Because p_m and p both assume exactly the same dimensionality of $\mathbf{r}$ (and $\mathbf{s}$), the only reasonable mapping of these two different sampling phases is onto one neuron with two different *modes* of activity. In Figure S2c, we show that one possible biological mapping is to propose that feedforward inputs (active when $\gamma = 1$) to the

basal dendrites of pyramidal neurons allow neurons to sample from p, and top-down inputs (active when $\gamma = 0$) to the apical dendrites of pyramidal neurons allows neurons to sample from p_m: interestingly, a corollary of this mapping is that a network could achieve 'detachability' by manipulating γ to generate sample network states in the absence of stimuli.

It is important to note that several normative plasticity models have proposed that top-down signals to the apical dendrites could serve as some form of training signal. We will adopt a similar attitude, and now assess the locality of the Wake-Sleep parameter updates with respect to this model formulation. If we take the sample size for our updates to be $K = 1$, based on Eqs. 53 and 57, for a single pair of samples $\mathbf{r}$, $\mathbf{s}$, we have:

$$\Delta \mathbf{W}^{in}_{ij} \propto \frac{1-\gamma}{\sigma_r^2} \left[(\mathbf{r}_i - f_i(\mathbf{W}^{in}\mathbf{s})) f_i'(\mathbf{W}^{in}\mathbf{s}) \mathbf{s}_j \right] \tag{60}$$

$$\Delta \mathbf{W}^{out}_{ij} \propto \frac{\gamma}{\sigma_s^2} \left[(\mathbf{s}_i - g_i(\mathbf{W}^{out}\mathbf{r})) g_i'(\mathbf{W}^{out}\mathbf{r}) \mathbf{r}_j \right]. \tag{61}$$

As with REINFORCE, both σ_r and σ_s are proportionality constants and can be disregarded. For $\Delta \mathbf{W}^{in}_{ij}$, a basal synapse on $\mathbf{r}_i$, several variables are required. First, the same signal that gates the influence of apical versus basal inputs, γ, must also *deactivate* plasticity at basal synapses. γ could be implemented in a neural circuit by either global inhibitory gating or by a neuromodulatory signal (Bredenberg et al., 2021)—whichever candidate signal would also have to gate plasticity. The synapse needs the postsynaptic firing rate $\mathbf{r}_i$, which is readily available, and a subtracted measure of current local to the basal compartment, $f_i(\mathbf{W}^{in}\mathbf{s})$—there is some indication that local dendritic voltage levels can affect synaptic plasticity, but the sign and exact form of this effect is variable across studies (Letzkus et al., 2006; Froemke et al., 2005; Sjöström and Häusser, 2006). As with REINFORCE, the synapse would require $f_i'(\mathbf{W}^{in}\mathbf{s})$, which is simply a monotonic function of $(\mathbf{W}^{in}\mathbf{s})_i$, and could be easily approximated; lastly, it would need the presynaptic firing rate $\mathbf{s}_j$. The information requirements for $\mathbf{W}^{out}_{ij}$ are almost exactly the same.

In terms of requiring only functions of pre- and postsynaptic activity, with the addition of some limited global context signal γ, these plasticity rules are plausibly local (Section 2.2). However, several features of this setup are unconfirmed, the most obviously testable being the Wake-Sleep sampling dynamics postulated by Eqs. 60 and 61: it seems unlikely that a neural network would entirely and synchronously switch into a 'generative' or hallucinatory regime for an extended period of time when $\gamma = 0$, and such a regime could not possibly occur in an awake, behaving animal, meaning that $\mathbf{W}^{in}$ could not be learned online (Section 2.5). However, a softer form of Wake-Sleep has been proposed (Bredenberg et al., 2021) which does allow for online learning, and does not interfere with active perception, suggesting that the principles established by Wake-Sleep may extend to more realistic formulations of γ. The strongest test (Section 2.7) of this family of algorithms is that artificially magnifying the influence of apical dendrites in a neural circuit should induce generative

sampling, i.e. hallucination; other models of apical dendritic learning (Sacramento et al., 2017; Guerguiev et al., 2017; Payeur et al., 2021; Urbanczik and Senn, 2014) do not propose this as a mechanism. Notice that this prediction requires our specific mapping of the Wake-Sleep algorithm onto neural circuitry: other interpretations are conceivable, and would have different predictions.

As we have discussed in Section D.1, the Wake-Sleep algorithm is capable of optimizing a broad range of *unsupervised* learning objectives, considerably more general than for instance Oja's rule (Oja, 1982) (though the specific toy example we provide is just a nonlinear form of probabilistic PCA). Unlike REINFORCE, the Wake-Sleep algorithm is unable to optimize reinforcement learning objectives, however, within the range of objectives that Wake-Sleep *can* optimize, it is typically much more scalable than REINFORCE (Section 2.6)[4]: in this way, it is an ideal complement, and having both algorithms or some hybridized form present in a neural circuit could be very powerful. However, the Wake-Sleep algorithm involves more approximation than REINFORCE. One could very easily wonder: since we have presented two sets of parameters in the Wake-Sleep algorithm minimizing two different objective functions, why should we expect the algorithm to converge or reliably improve performance on either objective?

To this point, we have identified two strange features of the Wake-Sleep algorithm that go hand-in-hand. First, it is strange that we should require a period of hallucinatory activity to train our parameters. Second, it is hard to interpret the convergence of an algorithm that is alternatively minimizing two slightly different objective functions: why all the work and extra conceptual baggage? Why not just do approximate gradient descent as we did with the REINFORCE algorithm and be done with it? In Section D.4.1 we will motivate why more standard gradient descent methods are not appropriate for this type of unsupervised learning, and in Section D.4.2 we will address the convergence properties of the Wake-Sleep algorithm from two different perspectives, explaining why the algorithm has such good empirical performance despite its approximations.

D.4.1 Why gradient descent with $\mathbf{W}^{in}$ won't work

Sometimes, to genuinely understand an algorithm, it's important to understand the weaknesses of alternative approaches. For didactic reasons, we will explore what happens if we simply take the gradient of $\mathcal{L}_{Wake}$ with respect to $\mathbf{W}^{in}$. We have:

[4]Though it still performs worse than backpropagation (Kingma and Welling, 2014; Rezende et al., 2014).

$$-\frac{d\mathcal{L}_{Wake}}{d\mathbf{W}_{ij}^{in}} = -\frac{d}{d\mathbf{W}_{ij}^{in}} \int \ln\left(\frac{p(\mathbf{r},\mathbf{s};\mathbf{W}^{in})}{p_m(\mathbf{r},\mathbf{s};\mathbf{W}^{out})}\right) p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) d\mathbf{r}d\mathbf{s} \tag{62}$$

$$= -\int \left[\frac{d}{d\mathbf{W}_{ij}^{in}} \ln\left(\frac{p(\mathbf{r},\mathbf{s};\mathbf{W}^{in})}{p_m(\mathbf{r},\mathbf{s};\mathbf{W}^{out})}\right)\right] p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) d\mathbf{r}d\mathbf{s}$$

$$-\int \ln\left(\frac{p(\mathbf{r},\mathbf{s};\mathbf{W}^{in})}{p_m(\mathbf{r},\mathbf{s};\mathbf{W}^{out})}\right) \frac{d}{d\mathbf{W}_{ij}^{in}} p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) d\mathbf{r}d\mathbf{s} \tag{63}$$

$$= -\int \left[\frac{d}{d\mathbf{W}_{ij}^{in}} \ln\left(p(\mathbf{r},\mathbf{s};\mathbf{W}^{in})\right)\right] p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) d\mathbf{r}d\mathbf{s}$$

$$-\int \ln\left(\frac{p(\mathbf{r},\mathbf{s};\mathbf{W}^{in})}{p_m(\mathbf{r},\mathbf{s};\mathbf{W}^{out})}\right) \frac{d}{d\mathbf{W}_{ij}^{in}} p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) d\mathbf{r}d\mathbf{s}, \tag{64}$$

where the second equality follows from the product rule, and the third equality follows from the fact that $\ln p_m(\mathbf{r},\mathbf{s};\mathbf{W}^{out})$ does not depend on $\mathbf{W}^{in}$. Interestingly, the first term in this equation is zero. To see this, we note the following sequence of identities:

$$\begin{aligned}\int \left[\frac{d}{d\mathbf{W}_{ij}^{in}} \ln\left(p(\mathbf{r},\mathbf{s};\mathbf{W}^{in})\right)\right] p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) d\mathbf{r}d\mathbf{s} &= \int \left[\frac{d}{d\mathbf{W}_{ij}^{in}} e^{\ln\left(p(\mathbf{r},\mathbf{s};\mathbf{W}^{in})\right)}\right] d\mathbf{r}d\mathbf{s} \\ &= \int \frac{d}{d\mathbf{W}_{ij}^{in}} p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) d\mathbf{r}d\mathbf{s} \\ &= \frac{d}{d\mathbf{W}_{ij}^{in}} \int p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) d\mathbf{r}d\mathbf{s} = \frac{d}{d\mathbf{W}_{ij}^{in}} 1 = 0.\end{aligned} \tag{65}$$

The first term is zero, which leaves only the second term of Eq. 64. It gives us:

$$-\frac{d\mathcal{L}_{Wake}}{d\mathbf{W}_{ij}^{in}} = -\int \ln\left(\frac{p(\mathbf{r},\mathbf{s};\mathbf{W}^{in})}{p_m(\mathbf{r},\mathbf{s};\mathbf{W}^{out})}\right) \frac{d}{d\mathbf{W}_{ij}^{in}} p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) d\mathbf{r}d\mathbf{s} \tag{66}$$

$$= \int \ln\left(\frac{p_m(\mathbf{r},\mathbf{s};\mathbf{W}^{out})}{p(\mathbf{r},\mathbf{s};\mathbf{W}^{in})}\right) \left(\frac{d}{d\mathbf{W}_{ij}^{in}} \ln p(\mathbf{r},\mathbf{s};\mathbf{W}^{in})\right) p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) d\mathbf{r}d\mathbf{s}, \tag{67}$$

where for the second equality we have once again used the identity in Eq. 65. Fascinatingly enough, this is exactly equivalent to the REINFORCE update (Eq. 21), if we take $R(\mathbf{r},\mathbf{s}) = \ln(p/p_m)$. Though the REINFORCE update might be practical for environmental rewards that an animal might receive, this particular choice of $R(\mathbf{r},\mathbf{s})$ requires detailed knowledge of the inner workings of a neural representation.

Not only is it not possible for an environmental signal to carry this information, there is no evidence that any neuromodulatory center in the brain is able to compute such a complicated signal based on neural network activity. Thus, even though this update appears to have the form of a reward-modulated Hebbian plasticity rule, there is very little reason to believe that it is local (Section 2.2). Furthermore, this form of update is well-known to have severe scalability (Section 2.6) issues, and demonstrably performs worse than Wake-Sleep on high-dimensional datasets (Werfel et al., 2003; Bredenberg et al., 2021). The Wake-Sleep algorithm is very much a response to these failings, using a local error signal specific to each neuron, rather than correlating each neuron's activity with a global reward signal. However, the Wake-Sleep algorithm employs more approximations than REINFORCE. In Section D.4.2, we will analyze the convergence properties of Wake-Sleep.

D.4.2 The convergence of Wake-Sleep

Currently, we have two updates that are approximating gradient descent on two different objectives: $\Delta \mathbf{W}_{ij}^{out} \approx -\lambda \frac{d\mathcal{L}_{Wake}}{d\mathbf{W}_{ij}^{out}}$, and $\Delta \mathbf{W}_{ij}^{in} \approx -\lambda \frac{d\mathcal{L}_{Sleep}}{d\mathbf{W}_{ij}^{in}}$, where λ is a small positive learning rate. In Section 2.1, we stressed the importance of viewing plasticity updates as decreasing a *unified* objective, but here we have two. How do we know that $\Delta \mathbf{W}_{ij}^{in}$ won't *increase* $\mathcal{L}_{Wake}$ and vice versa? Clearly, $\mathcal{L}_{Sleep}$ and $\mathcal{L}_{Wake}$ are closely related: one way of resolving this difficulty is by demonstrating that $\Delta \mathbf{W}_{ij}^{in} \approx -\lambda \frac{d\mathcal{L}_{Wake}}{d\mathbf{W}_{ij}^{in}}$. In this case, during the Wake phase, the system would optimize $\mathcal{L}_{Wake}$ with respect to $\mathbf{W}^{out}$, and during the Sleep phase, it would approximately optimize the same objective with respect to $\mathbf{W}^{in}$—this would amount to an approximation of coordinate descent. In fact, under certain conditions, it turns out that this is exactly what the Wake-Sleep algorithm is doing.

To see this, we begin with the REINFORCE-like update (Eq. 67) for gradient descent on $\mathcal{L}_{Wake}$:

$$-\frac{d\mathcal{L}_{Wake}}{d\mathbf{W}_{ij}^{in}} = \int \ln\left(\frac{p_m(\mathbf{r},\mathbf{s};\mathbf{W}^{out})}{p(\mathbf{r},\mathbf{s};\mathbf{W}^{in})}\right)\left(\frac{d}{d\mathbf{W}_{ij}^{in}} \ln p(\mathbf{r},\mathbf{s};\mathbf{W}^{in})\right) p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) d\mathbf{r} d\mathbf{s}. \tag{68}$$

Interestingly, we notice that if $p_m \approx p$, then by first-order Taylor expansion, $\ln(p_m/p) \approx p_m/p - 1$. Plugging this approximation in (see (Bredenberg et al., 2021) for a more detailed justification of this approximation), we get:

$$-\frac{d\mathcal{L}_{Wake}}{d\mathbf{W}_{ij}^{in}} \approx \int \left(\frac{p_m(\mathbf{r},\mathbf{s};\mathbf{W}^{out})}{p(\mathbf{r},\mathbf{s};\mathbf{W}^{in})} - 1\right)\left(\frac{d}{d\mathbf{W}_{ij}^{in}} \ln p(\mathbf{r},\mathbf{s};\mathbf{W}^{in})\right) p(\mathbf{r},\mathbf{s};\mathbf{W}^{in}) d\mathbf{r} d\mathbf{s} \tag{69}$$

$$= \int \left(\frac{d}{d\mathbf{W}_{ij}^{in}} \ln p(\mathbf{r},\mathbf{s};\mathbf{W}^{in})\right) p_m(\mathbf{r},\mathbf{s};\mathbf{W}^{out}) d\mathbf{r} d\mathbf{s} \tag{70}$$

$$= -\frac{d\mathcal{L}_{Sleep}}{d\mathbf{W}_{ij}^{in}}, \tag{71}$$

where for the first equality we have once again used the identity Eq. 65. Essentially, if a global optimum such that $p_m = p$ exists, it is shared by both $\mathcal{L}_{Wake}$ and $\mathcal{L}_{Sleep}$. Thus, we can expect the gradients of these two objective functions to behave very similarly if p_m is close to p. Because the Wake phase (updating $\mathbf{W}^{out}$) occurs without approximation, the algorithm has the opportunity to enter this regime before the approximating Sleep phase ever occurs.

An alternative analysis of the Wake-Sleep algorithm (Dayan et al., 1995) observes that for fixed $\mathbf{W}^{out}$, $\mathcal{L}_{Sleep}$ and $\mathcal{L}_{Wake}$ share a global minimum with respect to $\mathbf{W}^{in}$ when $p_m(\mathbf{r}|\mathbf{s};\mathbf{W}^{out}) = p(\mathbf{r}|\mathbf{s};\mathbf{W}^{in})$, as long as there exists a $\mathbf{W}_{opt}^{in}$ such that this equality holds. If $\mathcal{L}_{Sleep}$ is convex and this global minimum is attainable, fully optimizing $\mathcal{L}_{Sleep}$ with respect to $\mathbf{W}^{in}$ during the Sleep phase is therefore guaranteed to also optimize $\mathcal{L}_{Wake}$. Therefore, as long as these to conditions of convexity and attainability of the global minimum are satisfied (they are not in general, but do hold for simple examples like Factor Analysis (Amari and Nakahara, 1999)), both phases decrease $\mathcal{L}_{Wake}$. Rather than an approximation of coordinate descent, this can be viewed as an approximation of the Expectation-Maximization (EM) algorithm (Dempster et al., 1977).

We see that there are two different ways of interpreting Wake-Sleep: first, it is an approximation of coordinate descent that becomes a better approximation the closer to the optimum it becomes. Second, under restricted conditions, Wake-Sleep can be viewed as an approximation of the EM algorithm. Both of these perspectives are conditional on assumptions about the probability models being trained, requiring a generative model $p_m(\mathbf{r},\mathbf{s})$ and a forward map $p(\mathbf{r}|\mathbf{s})$ capable of mutually reaching good performance for an environmental stimulus distribution $p(\mathbf{s})$. Though Wake-Sleep empirically performs quite well under a variety of stimulus conditions and network models (Dayan and Hinton, 1996), these are important caveats: the comparative weakness of the demonstrations of Wake-Sleep's convergence relative to gradient descent or EM is a common point of criticism of the algorithm (Rezende et al., 2014; Kingma and Welling, 2014; Mnih and Gregor, 2014).