Biol Cybern (2009) 100:319–330
DOI 10.1007/s00422-009-0305-x

ORIGINAL PAPER

Biological Cybernetics

Stimulus sampling as an exploration mechanism for fast reinforcement learning

Boris B. Vladimirskiy · Eleni Vasilaki · Robert Urbanczik · Walter Senn

Received: 5 March 2009 / Accepted: 18 March 2009 / Published online: 10 April 2009

Abstract Reinforcement learning in neural networks requires a mechanism for exploring new network states in response to a single, nonspecific reward signal. Existing models have introduced synaptic or neuronal noise to drive this exploration. However, those types of noise tend to almost average out—precluding or significantly hindering learning—when coding in neuronal populations or by mean firing rates is considered. Furthermore, careful tuning is required to find the elusive balance between the often conflicting demands of speed and reliability of learning. Here we show that there is in fact no need to rely on intrinsic noise. Instead, ongoing synaptic plasticity triggered by the naturally occurring online sampling of a stimulus out of an entire stimulus set produces enough fluctuations in the synaptic efficacies for successful learning. By combining stimulus sampling with reward attenuation, we demonstrate that a simple Hebbian-like learning rule yields the performance that is very close to that of primates on visuomotor association tasks. In contrast, learning rules based on intrinsic noise (node and weight perturbation) are markedly slower. Furthermore, the performance advantage of our approach persists for more complex tasks and network architectures. We suggest that stimulus sampling and reward attenuation are two key components of a framework by which any single-cell supervised learning rule can be converted into a reinforcement learning rule for networks without requiring any intrinsic noise source.

Keywords Online learning · Hebbian learning · Association task—Noise · Reward · Punishment · Reward attenuation · Hippocampus · Medial temporal lobe · Striatum

1 Introduction

Reinforcement learning is a process whereby stimulus-response behavior is adjusted to maximize a single reward signal. This type of learning is involved in the acquisition of associative memories, i.e., long-term associations between a stimulus and a desired response. Associative learning has been extensively studied in the rat and monkey in many brain areas, such as the medial temporal lobe and the hippocampus in particular (Cahusac et al. 1993; Wirth et al. 2003; Suzuki 2007), motor regions of the frontal lobe (Mitz et al. 1991; Chen and Wise 1995a,b; Brasted and Wise 2004), prefrontal cortex (Asaad et al. 1998), and the striatum (Schultz et al. 2003; Brasted and Wise 2004; Pasupathy and Miller 2005; Williams and Eskandar 2006), on a variety of tasks, commonly including conditional motor associative learning in which sensory stimuli have to be associated with correct motor responses. Specifically, the existence of prominent subpopulations of cells selectively altering their firing rates during the course of learning in correlation with the behavioral performance has been established (Suzuki 2007). Long-term potentiation (LTP) and long-term depression (LTD) are assumed to underlie this and other kinds of long-term learning at the level of individual synapses. However, mechanisms that drive the synaptic plasticity in order to enable the fast learning of random new associations, for which the hippocampus is critically important (Vargha-Khadem et al. 1997;

This work was supported by the Swiss National Science Foundation grant K-32K0-118084.

B. B. Vladimirskiy (✉) · R. Urbanczik · W. Senn
Department of Physiology, University of Bern,
Bühlplatz 5, 3012 Bern, Switzerland
e-mail: vladimirski@pyl.unibe.ch; bbv201@nyu.edu

E. Vasilaki
Laboratory of Computational Neuroscience (IC/LCN),
Ecole Polytechnique Fédérale de Lausanne, Station 15,
1015 Lausanne, Switzerland

Bayley and Squire 2002; Stark et al. 2002; Stark and Squire 2003), are unknown. Specifically, what are the mechanisms allowing the neural network to find the synaptic efficacies that lead to fast progress in learning? Are these mechanisms synaptic, neuronal, network-level or extrinsic? Addressing these questions directly in an experiment is problematic, thus making a modeling approach desirable.

Making a useful interpretation of the global reward signal at every synapse in a large network is, however, a difficult modeling problem as well. Hence, trial and error is usually employed to explore the synaptic efficacies. In general, any neural network model of reinforcement learning needs a mechanism for generating altered responses and a mechanism to incorporate the external feedback. While external feedback is typically assumed to be mediated by a global neurotransmitter signal, especially dopamine, modulating synaptic plasticity (Daw and Doya 2006; Wickens et al. 2007), little is known about the mechanism generating the response variability. Different noise sources underlying the response exploration have been suggested, ranging from stochastic neuronal activation (Barto and Jordan 1987; Williams 1992) to noisy current injection at the soma (Xie and Seung 2004) or the synapse (Doya and Sejnowski 1998; Fiete and Seung 2006), to stochastic neurotransmitter release (Seung 2003), probabilistic winner-take-all strategy (Vasilaki et al. 2009), and stochastic reward delivery (Montague et al. 1995). Although the introduction of an explicit noise source in these models simplifies the mathematical treatment, the inherent stochasticity often leads to the excessive exploration of network states and thus makes many more biologically plausible learning rules slow (Werfel et al. 2005). Furthermore, all noise-based learning rules suffer from the need to tune the amount of noise applied carefully: too much noise yields unreliable performance, and too little, no (or very slow) exploration.

Stochastic synaptic or neuronal processing as a possible exploration mechanism becomes even more problematic when the information is encoded in population firing rates or mean firing rates, as is believed to be the case in higher cortical areas (Aggelopoulos et al. 2005; Rolls et al. 2006). In fact, spatial (e.g., in a multilayered network) or temporal averaging flattens any independent noise which is originally present at the level of the single spiking neuron or synapse. Hence, the desired variability in the firing rates is severely compromised. Therefore, unless the individual noise is correlated (which requires a corresponding neural architecture), other exploration mechanisms must be considered in the context of reinforcement learning based on mean population or temporal firing rates.

Here we argue that online learning in a complex stimulus environment solves the exploration problem by itself. Ongoing synaptic plasticity triggered by the naturally occurring online sampling of a stimulus out of an entire stimulus set produces enough response variability to effectively explore new network states. This implicit noise source (we call it stimulus sampling) is self-regulating and does not require additional tuning. While the most stochastic stimulus sampling results if the order in which the stimuli are sampled is random, it is sampling per se, rather than its random or not order, that plays the main role in generating this implicit noise. Stimulus sampling is assumed in this work to be outside of learner's control, which is the case in many biologically important learning tasks, such as learning to distinguish food from non-edible objects early in life.

The second key component of our approach is learning from the difference between the reward signal and its expected value (we call it reward attenuation). Reward attenuation is biologically-inspired (Schultz 2002; Bayer and Glimcher 2005; Daw and Doya 2006; Kobayashi and Okada 2007; Schönberg et al. 2007) and allows effective learning by maximizing the exploration at the start of the learning process and minimizing it in response to rewarded trials only (Bayer and Glimcher 2005) at the end, so that no repeated learning and unlearning of the same stimuli occur.

In order to explore the feasibility of stimulus sampling combined with reward attenuation, we consider Hebbian reinforcement learning (HRL) as a simple, biologically plausible generalization of perceptron Hebbian learning (Hebb 1949; Rosenblatt 1958; the error-correcting nature of HRL is also similar to that of the model by Rescorla and Wagner (1972)) to learning from both reward and punishment in multilayered networks and/or with multiple output units. When learning from punishment only, our learning rule can be considered a form of the associative reward-penalty (ARP) algorithm (Barto and Jordan 1987; Hertz et al. 1991; Williams 1992) in the limit of small stochasticity in the response function of the single neuron. Both HRL and ARP address the problem of spatial credit assignment among all synapses in a multilayered network that is to implement stimulus-reward associations. In contrast, the theory of temporal-difference (TD) learning, while bearing some analogy to our approach in the special case of reward delivery immediately following network's response to a stimulus (Montague et al. 1996), generally addresses the temporal credit assignment problem for state-action sequences (Sutton and Barto 1981, 1998; Schultz et al. 1997; McClure et al. 2003; Seymour et al. 2004; Doya 2008) without offering a simple neural network implementation. Furthermore, the focus of the previous research has not been on comparing the effectiveness of different stochasticity sources for the exploration purposes, which, rather than the advantages of HRL as such, is the main emphasis of this paper.

We first show that the fast performance of HRL is in excellent correspondence to that of macaque monkeys on stimulus-response association tasks (Chen and Wise 1995a; Wirth et al. 2003). In contrast, reinforcement learning rules based on

node perturbation (NP) and weight perturbation (WP) are too slow to explain the data: their intrinsic noise interferes with stimulus sampling and both of those rules are not Hebbian. The superiority in convergence speed and reliability of HRL becomes even more pronounced when more difficult problems and networks with hidden layers are considered. To prove that the high performance of HRL originates precisely from stimulus sampling, we compare batch learning with our normal online learning: when synaptic updates are only administered at the end of an epoch containing all stimuli, instead of after each stimulus presentation, learning fails.

Furthermore, we demonstrate that learning from mistakes only is generally slower than combined learning from punishment and reward, questioning earlier statements that learning from mistakes is the preferred strategy (Chialvo and Bak 1999). Learning in the case of rewarded events becomes necessary if the number of output units is large, making rewarded responses initially rare. Discarding the information provided by these rare events, hence, leads to drastic deterioration in performance.

Finally, we suggest that combining stimulus sampling with reward attenuation represents a general framework by which any supervised learning rule for a single neuron can be turned into the corresponding reinforcement learning rule for the whole network independent of the modalities involved.

2 Results

2.1 Comparison to primates' performance

We first studied how well our model's performance corresponds to that of primates on visuomotor association tasks (Chen and Wise 1995a; Wirth et al. 2003). Following foveal fixation at the center of a computer screen, the monkey was shown a complex natural (Wirth et al. 2003) or artificial (Chen and Wise 1995a) scene superimposed with four saccade targets at the top, bottom, and the sides of the screen (Fig. 1a). The goal was to learn the correct associations of each scene with one of the four saccade targets (randomly chosen for each scene and fixed throughout the experiment). All scenes were equally likely to appear and were presented in randomized order. Randomly intermixed with one to four novel scenes, the monkey was also shown two to four very familiar stimuli whose associations it had learned previously.

In order to match the experimental protocol, we used a simple network with two binary outputs representing the four possible saccade responses (Fig. 1b). The input layer coding for natural scenes in the experiments consisted of a large number of binary neurons. Simulated stimuli were random and thus fulfilled the experimental requirement that the natural scenes be very different. Eight associations were generated randomly, the network was allowed to learn four of those first, starting from a random synaptic configuration, and then had to learn the full set starting with the synaptic weights achieved at the end of the first learning phase.

In Fig. 1c, the performance (instantaneous reward averaged over 1000 independent learning sessions) on this learning task of three representative learning rules (cf. Sect. 4), Hebbian reinforcement learning (HRL), node perturbation (NP) and weight perturbation (WP), is shown. For each rule, the synaptic efficacy change is proportional to the presynaptic signal and attenuated reward (cf. Sect. 2.3). The other key factor to which the change is proportional varies among the three rules. For HRL, this factor is equal to the difference between the postsynaptic signal and the excitation/inhibition balance point of 0.5; for NP, the difference is the noise signal applied to the postsynaptic neuron, whereas for WP, the difference is the noise applied to the synapse. The performance of HRL is in excellent correspondence to that of the monkey: both the monkey and HRL require approximately 11–12 presentations per novel stimulus until the full set of stimulus-response associations is reproduced correctly (Fig. 1d). In contrast, NP and WP are too slow to capture the monkey's performance; NP, which was the faster of the two, produced the median learning time of approximately 28 trials per new association.

An interesting additional experimental observation was that during learning the monkey made almost no mistakes on the familiar associations (2% reported by Chen and Wise (1995a)). To test the models in that respect, we computed the mean percentage of errors on the familiar associations and found the values of 2.4% for HRL, 4.7% for NP, and 11.8% for WP. Hence, HRL is closest to the data. The high percentages of errors on previously learned associations for NP and WP are due to the very nature of these learning rules which require a fair amount of intrinsic noise for faster initial convergence. However, at a later stage of learning this amount of intrinsic noise combined with the non-Hebbian character of NP and WP results in more re-learning of the previously learned stimuli than in the case of the HRL, which is Hebbian and not intrinsic noise-based. Lowering the amount of noise does result in less re-learning (data not shown), but the performance suffers significantly.

2.2 Stimulus sampling is essential for learning

We next demonstrate that the use of stimulus sampling induces effective noise that is critical to the sufficient exploration of the synaptic weight space and, hence, to successful learning. To isolate the effect of the ongoing synaptic modifications we contrast our usual online learning with batch learning for the HRL rule on the same monkey problem as in Fig. 1. In the batch learning scenario, the computed synaptic weight changes are not applied instantaneously after each stimulus presentation (as they are in online learning),

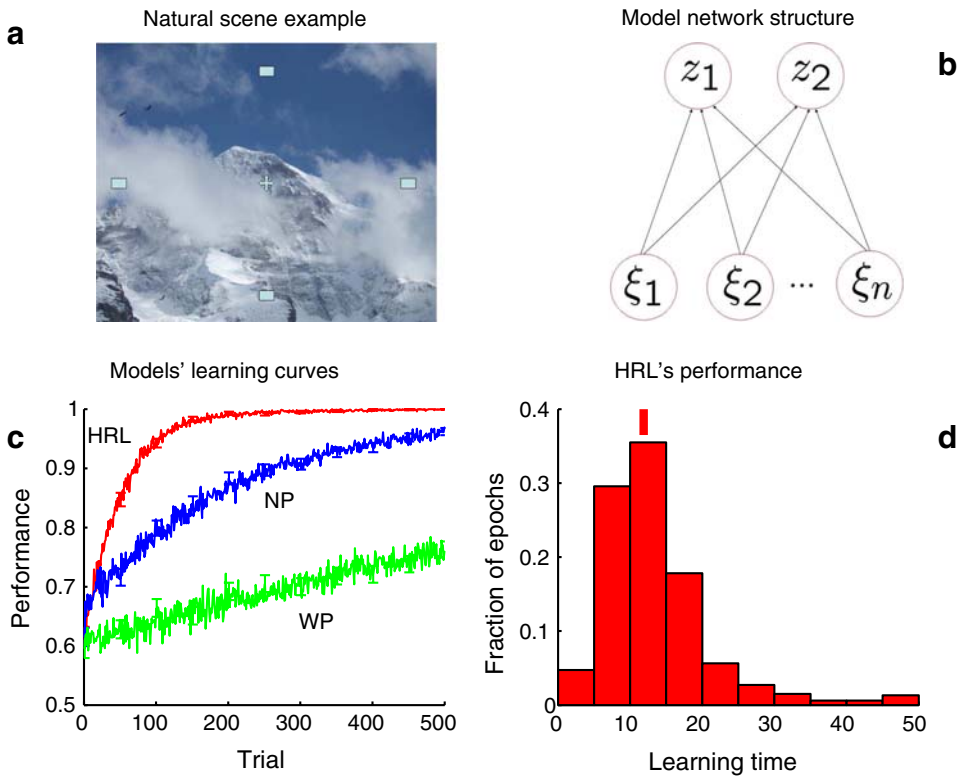

Fig. 1 Our model's performance is in good correspondence to the monkey data by Chen and Wise (1995a) and Wirth et al. (2003). **a** A possible natural scene used by Wirth et al. (2003) is shown (Chen and Wise (1995a) used complex artificial visual stimuli); superimposed are the fixation spot and four saccade targets. The task is to classify correctly (make a saccade to the correct target for the monkey) four novel stimuli having learned the correct responses to another four. **b** Model network structure: two binary output neurons code for the four possible targets; all synaptic weights are analog, bounded and subject to global inhibition (Sect. 4.1). The number of afferents n equals 1000. **c** Instantaneous performance curves (with each half of an error bar indicating one standard error of the mean) are shown, for our HRL rule (*red*), node perturbation (NP, *blue*) and weight perturbation (WP, *green*) learning rules. A point on a performance curve is the percentage of correct responses on the corresponding trial averaged over 1000 random, independently learned input–output associations and random initial synaptic weights. Parameter values leading to the fastest convergence were used for each model. **d** Histogram of learning times for HRL. The performance is in excellent correspondence to that of the monkey: HRL produces the median learning time (number of trials per new stimulus until all stimuli have been learned, indicated by the thin red bar at the top) of 12 and the mean of all learning times less than two magnitudes of the median of 11.7 ± 0.16 versus 10.6 ± 3.4 and 12 ± 1 for the monkey reported by Chen and Wise (1995a) and Wirth et al. (2003), respectively. Both NP and WP are too slow to capture the monkey's performance. The results do not vary by more than 10% with n decreasing to 100 or increasing beyond 1000

but are accumulated instead and applied after each epoch of P trials, where P is the number of different associations to be learned ($P = 8$ in our case). In what follows, we contrast random-order batch learning, in which stimuli are presented in a random order, with fixed-order batch learning in which the stimuli are always presented in the same order. The order of stimuli presentation within a batch, however, is not important as long as all the stimuli appear exactly once; due to the nature of the batch synaptic update, the results would be exactly the same.

In fixed-order batch learning, the noise induced by stimulus sampling is suppressed entirely. The comparison of batch and online learning shows that the performance of fixed-order batch is very poor indeed, just slightly above the chance level (the cyan curve in Fig. 2a vs. the red curve). When applied to the same network state, fixed-order batch learning always results in the same synaptic modifications; thus, the exploration becomes too limited. In contrast, random-order batch learning retains some noise since stimuli presented between the updates are not always exactly the same even though, on average, each stimulus is presented once per update. Due to the remaining stochasticity, random-order batch is significantly better but still far worse than the online HRL with random stimulus sampling (cf. golden and red curves, respectively, in Fig. 2a) since the amount of effective noise is greatly reduced in the former case. This decrease in noise also accounts for the widening of the histogram of learning times in the case of random-order batch, while for online stimulus sampling learning is reliably fast (Fig. 2b). We conclude that even in the case of learning only eight associations, the exploration of synaptic efficacies induced by the random order of stimulus presentation is sufficient to

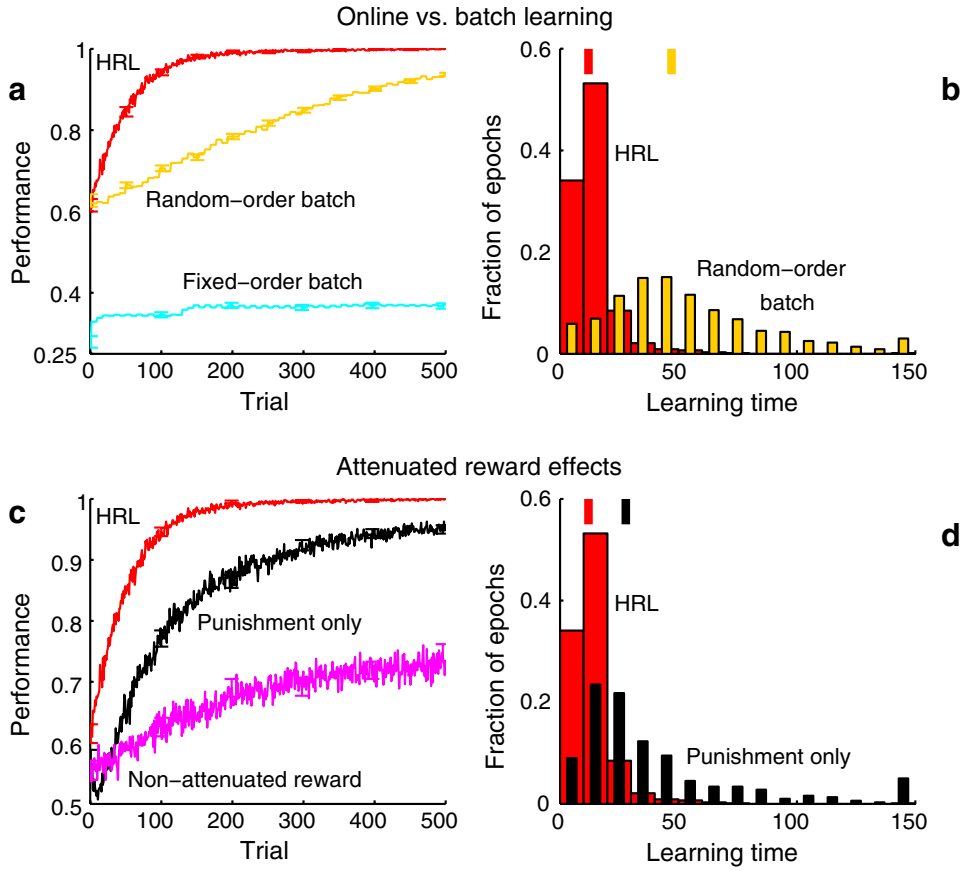

Fig. 2 Stimulus sampling and attenuated reward are essential for learning. **a** The online learning performance curve of HRL (*red*, the same as in Fig. 1) and the batch learning curves for random (*golden*) and fixed (*cyan*) stimulus presentation order, also for HRL. For the fixed-order batch learning no noise is effectively generated, resulting in extremely limited exploration and the performance is just slightly above the chance level. Random-order batch, however, allows for restricted exploration which eventually leads to learning of the task, albeit significantly slower than for the online HRL. **b** Corresponding distributions of learning times for HRL and random-order batch are shown in the same colors as in **a**. **c** Learning from mistakes only (*black curve*) is poor as compared to learning from reward and mistakes (*red curve*, same as in **a**) for the case of multiple output units (e.g., for the monkey association task). In both cases online learning with reward attenuation was used. Unattenuated online learning from reward and punishment destabilizes the learning process as it progresses (*magenta*). **d** Corresponding distributions of learning times for HRL and learning from punishment only are shown in the same colors as in **c**. Optimum parameters were used for each curve/histogram and the monkey task was the same as in Fig. 1

efficiently explore the synaptic state space, resulting in fast and robust learning.

2.3 Reward attenuation is necessary for achieving the maximum reward

Different forms of combining the reward signal with synaptic modifications are possible. In HRL these modifications are Hebbian and attenuated on rewarded trials (cf. Bayer and Glimcher 2005 and Sect. 4.2) so that the further the learning process progresses (i.e., the closer the average reward approaches 1), the smaller the synaptic changes are. On the other hand, on non-rewarded trials anti-Hebbian modifications are applied without any attenuation based on the reward history (cf. Bayer and Glimcher 2005). Both the synaptic changes in the case of reward and their attenuation are essential. To reveal the benefits of these two ingredients we again consider the previously described monkey problem.

When only the anti-Hebbian component on non-rewarded trials was applied, learning slowed down considerably (the red curve in Fig. 2c vs. the black curve). Note that this form of learning from mistakes corresponds to applying the classical perceptron error correction rule (Rosenblatt 1958; Hertz et al. 1991) to each output neuron in the network. However, in contrast to the single output neuron situation, the Hebbian component becomes important in the presence of multiple output units. The more output units there are in the network, the more significant role the Hebbian component would play and the worse the perceptron rule would perform: the chance of producing a correct response out of the many possibilities is poor at the initial stage of learning because the initial synaptic weights are random. The information provided by

rare rewarded trials is therefore very valuable, necessitating learning from reward in general.

At the late stage of learning, on the other hand, the attenuation of synaptic changes on rewarded trials becomes important. If the synaptic changes do not attenuate with increasing average reward, the performance saturates early (Fig. 2b, magenta curve). This arises because strengthening the synaptic efficacies as strongly as during the early learning stage leads to interference among the representations of the already learned stimuli. Thus, learning does not stabilize as the same stimuli are learned and unlearned repeatedly. The histogram of learning times confirms that without reward attenuation there are always some associations that are not learned correctly. Reward attenuation also takes into account that, once the performance is good, an error is much more informative for further improvement than success.

2.4 Challenging problem in a network without hidden layers

To further explore the benefits of the stimulus sampling approach, we studied the performance of our model on difficult problems. First, we considered a challenging task in the network without hidden layers, similar to that depicted in Fig. 1b, but with 100 inputs and a single output neuron. The problem was to correctly classify 130 random stimuli into 2 randomly chosen, equiprobable classes. This problem is difficult since the number of stimuli exceeds that of the inputs (Hertz et al. 1991). The results are presented in Fig. 3a and b using the same colors and line styles as in Fig. 1. For optimum parameters, HRL is almost six times faster than NP with the median convergence times of 85 and 483, respectively.

WP does not converge at all beyond the chance level within 3000 presentations of each stimulus. Apparently, the noise level for this problem has to be very low in order to avoid the undesirable interplay among synaptic efficacy changes for different stimuli. However, such small noise levels necessarily result in exceedingly long learning times. This also applies to the observed fraction of non-convergent learning sessions for NP, which exceeds 10% (Fig. 3b). HRL, on the other hand, does not suffer from this problem as it does not rely on any internal noise source. In fact, the percentage of non-convergent learning sessions for HRL is negligible (Fig. 3b).

To explain the superior performance of HRL as compared to NP, we consider three representative synaptic time courses during a single learning session (Fig. 3c). Both learning rules are characterized by an initial learning phase during which synaptic efficacies tend to approach the non-biased value of 0.5, at which a synaptic efficacy is equidistant from both synaptic strength limits and is balanced by the global inhibition (Senn and Fusi 2005). This initial phase is followed by a second phase in which the exploration continues in order for the network to garner the maximum reward. In the second phase the asymptotic values are approached. For HRL the asymptotic values are approached at a nearly constant speed (disregarding the jitter) and reached within finite time. But for NP the second phase is significantly delayed compared to HRL: explicit noise causes jitter and less reliable performance. Furthermore, as for any gradient procedure with a smooth cost function, the convergence of NP slows down as learning progresses (the attractor is approached) and the asymptotic value is not reached in finite time. In other words, smaller and smaller noise would be required to prevent repeated learning and unlearning of the same stimuli as learning progresses with NP. The reward within the single learning session reflects the strong jitter in the case of NP whereas in the case of HRL the jitter decreases and eventually vanishes (Fig. 3d).

2.5 Stimulus sampling retains the performance advantage in networks with hidden layers

Reinforcement learning is particularly difficult for networks with more than one output unit or with hidden layers as a useful interpretation of the single reward signal becomes highly nontrivial. To demonstrate the advantages of stimulus sampling in the latter case, we considered neural networks with one (Fig. 4a1–c1), two (Fig. 4a2–c2), or three (Fig. 4a3–c3) hidden layers. The task was to correctly classify 20 random stimuli into 2 random classes, with each network possessing 5 inputs, 5 neurons in each hidden layer, and 1 output unit (Fig. 4c1–c3).

This problem cannot be solved without hidden layers as the number of stimuli exceeds the number of input neurons fourfold, which is well beyond the learning capacity of a single perceptron with five inputs (Hertz et al. 1991, also explicitly illustrated by the cyan curve in Fig. 4a1). With one hidden layer and a total of 25 synapses learning becomes possible, with approximately 0.8 stimuli to be stored per synapse. HRL required the median learning time of 232 presentations per stimulus (Fig. 4b1). When additional hidden layers were added, HRL's performance did not deteriorate by more than one-eigth (median learning times of 260 and 253 for two and three hidden layers, respectively, cf. Fig. 4b1–b3), although the synaptic state space significantly increased in dimensionality.

NP and WP were also able to learn the task in all three network configurations, but with, at best, threefold latency (for NP with one hidden layer) compared to HRL. Interestingly, the strong performance advantage that NP demonstrated compared to WP on problems without hidden layers (Figs. 1, 3) does not extend to networks with at least one hidden layer (Fig. 4b1–b3). The previously more targeted exploration in the case of NP with the efficacies of all presynaptic synapses onto a single neuron changed in concert is apparently not more effective than the completely independent

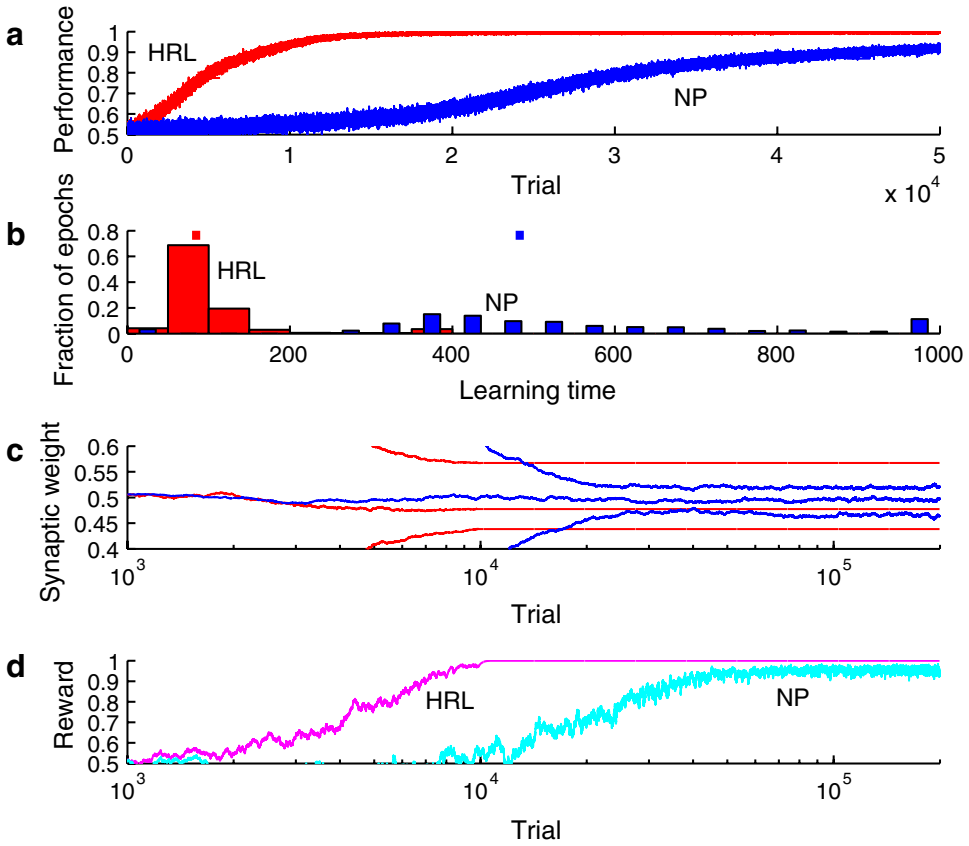

Fig. 3 HRL with stimulus sampling significantly outperforms NP and WP on a challenging two-class problem in a network with 100 input neurons and 1 output neuron. The task is to classify 130 random stimuli into those two random classes. **a** Learning curves for optimum parameters for HRL (*red*) and NP (*blue*), averaged across 1000 learning sessions. WP does not converge at all beyond the chance level. **b** Histograms of learning times. HRL is almost six times faster than the NP. Additionally, in more than 10% of all cases, NP does not converge (rightmost *blue bar* in **b**), whereas for HRL the fraction of non-convergent cases is negligible. The quantities shown are the same as those in Fig. 1. The median convergence times are 85 for HRL, and 483 for NP (*red* and *blue small squares*, respectively). **c** Synaptic weight time courses of three representative synapses (with initially strong, intermediate and low synaptic efficacies, respectively) for HRL (*red*) and NP (*blue*) show how the synaptic modifications for HRL stop in finite time once the maximum reward (**d**) has been garnered, as opposed to the slow asymptotic convergence for NP to a close neighborhood of 0.5. The network was initialized in the same state for both HRL and NP, and the sequence of stimulus presentations was the same during learning. **d** Instantaneous reward time courses corresponding to the synaptic time courses in panel **c** show a uniform decrease in jitter amplitude for HRL as learning progresses until it eventually vanishes when all the associations have been learned correctly (*red*). In contrast, the jitter amplitude remains virtually constant for NP throughout the whole learning period (*blue*) and the maximum reward is never gained

exploration of synaptic weights in networks with hidden layers.

Besides the longer learning time, the percentage of nonconvergent sessions for NP and WP was between 20 and 30%, rendering NP and WP essentially unusable for biological applications in this case. Note here that all the three learning rules were optimized for fastest median learning times. Smaller percentages of nonconvergent sessions could be achieved with NP and WP (data not shown), but at the expense of significant increases in the learning times (say, by a factor of 2). In contrast, HRL demonstrated the nonconvergent percentage of well below 10% without any additional tuning.

3 Discussion

In this study, we have shown that successful reinforcement learning in a neural network does not necessarily need to rely on an explicit noise source as has been assumed previously. We propose stimulus sampling as an alternative to the ongoing controversy on the adequate noise source in reinforcement learning rules (Seung 2003; Werfel et al. 2005; Fiete and Seung 2006). Stimulus sampling, based on the naturally occurring online sampling of a stimulus out of an entire stimulus set, provides enough fluctuations in the synaptic efficacies for the effective exploration of suitable stimulus-response associations. We applied stimulus sampling in the

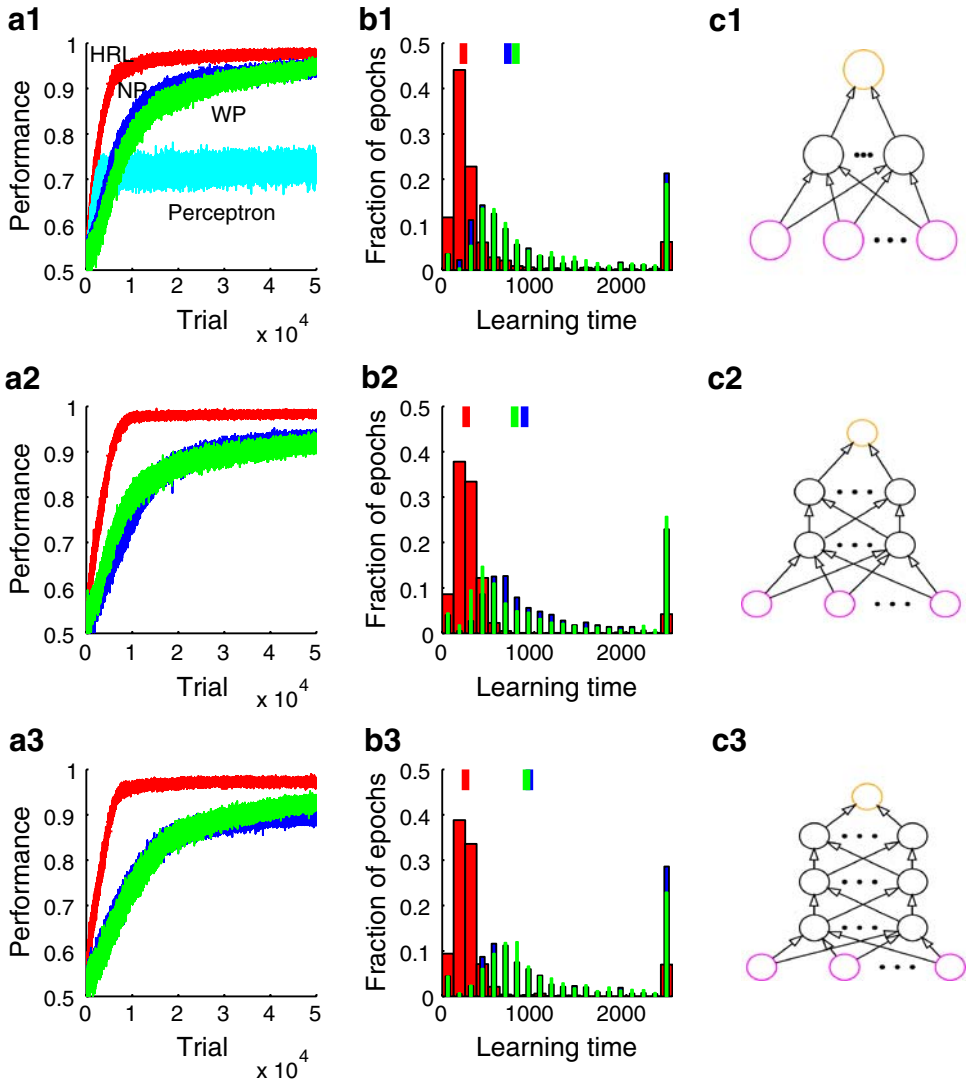

Fig. 4 HRL outperforms NP and WP in a network with hidden layers. The task is to classify 20 random stimuli into 2 random classes, with 5 input units and one, two, and three hidden layers, respectively (**c1**–**c3**). The problem genuinely requires hidden layers as revealed by the early saturation of learning with the perceptron (*cyan curve* in **a1**). HRL learns at least three times faster than NP and WP, with the learning times that are essentially independent of the number of layers (cf. *red curves* in **a1**–**a3** and *histograms* in **b1**–**b3**). Median convergence time for HRL (*red*), NP (*blue*) and WP (*green*) are, respectively, 232, 714 and 801 for one hidden layer (**b1**); 260, 896 and 788 for two hidden layers (**b2**); 253, 947 and 917 for three hidden layers (**b3**). Parameters resulting in the fastest learning were used in all cases

context of a Hebbian reinforcement learning rule, a simplest form of Hebbian plasticity which incorporates correlation-based LTP and anti-correlation-based LTD, modulated by an attenuating reward signal that dynamically decreases the amplitude of synaptic modifications on rewarded trials as learning progresses. We have shown that fast reinforcement learning in case of multiple output neurons requires synaptic modifications to be induced by rewarded trials, restricting the previously suggested learning from mistakes only (Chialvo and Bak 1999) to specific network architectures. Learning from reward is especially important during the initial learning phase when rewarded network states are rare and thus very informative. As learning continues, on the other hand, the amplitude of Hebbian learning modifications induced by reward must necessarily be attenuated to prevent undesirable interference in responses to different stimuli and achieve high performance.

We have found that the effectively stochastic presentation of stimuli due to the sampling for both simple and more complex networks and learning tasks results in the exploration levels that make HRL always superior to two well-established learning rules, node perturbation and weight perturbation, which are based on intrinsic neuronal and synaptic noise, respectively. The degree of exploration induced by stimulus sampling is automatically matched to the size and diversity of the stimulus set to be learned, and thus scales with the complexity of the learning task. For example, adult animals consciously attempt to improve their survival chances by seeking food and avoiding predators. Thus, they typically explore only a part of the entire stimulus space, which decreases the amount of the effective noise induced by stimulus sampling, but since the complexity of the learning problem is also reduced, less noise is necessary for the exploration. Similarly, adding temporal correlations by fixing

the order in which stimuli are presented, for instance, does not substantially hamper the learning performance as long as the synaptic updates are performed online and the stimulus set is not too small, guaranteeing sufficient exploration (learning curves are almost identical to those for HRL in all figures if stimuli are presented in the same order as for the fixed-order batch). In contrast, intrinsic noise sources cannot endow such self-regulating properties on the exploration process.

By investigating the evolution of jitter in the individual synaptic efficacy and reward curves during learning, we have demonstrated that stimulus sampling does in fact dynamically regulate the exploration as opposed to the noise-based learning rules, which maintain a constant level of jitter throughout the learning process. This accounts for the significantly faster and more reliable learning using HRL, as compared to NP and WP, on the visuomotor association task motivated by the monkey data (Chen and Wise 1995a; Wirth et al. 2003). In fact, only HRL corresponds very well to the learning speed of the animals. The superior performance of HRL persists for more challenging problems close to the learning capacity, and for problems involving feedforward networks with multiple layers of neurons. The fast performance of HRL could imply that stimulus sampling, unlike intrinsic noise, is especially relevant for the flexible fast learning of new associations that has been shown to be critically dependent on the intact medial temporal lobe and hippocampus (Vargha-Khadem et al. 1997; Bayley and Squire 2002; Stark et al. 2002; Stark and Squire 2003).

We have demonstrated that stimulus sampling represents a viable concept for learning stimulus-reward associations in a complex neuronal network by means of a single binary reinforcement signal, even with the simple HRL. Are there noise-based rules that could, unlike NP and WP, compete with it and, in particular, reproduce the primate's performance? One commonly used family of stochastic learning rules is associative reward-penalty (Barto and Jordan 1987; Hertz et al. 1991; Williams 1992). However, it has been shown previously that it can only compete with HRL if the intrinsic noise level is low (Vasilaki et al. 2009), in which case ARP approaches our HRL rule in the deterministic limit of learning from mistakes only. But we have shown that learning from mistakes only is already suboptimal in the case of only two output units (Fig. 2c, d) and would become exponentially worse as the number of outputs increases. This is not in contradiction to the results of Mazzoni et al. (1991) who had found that ARP could match the performance of back-propagation (Rumelhart et al. 1996), a theoretically attractive but biologically unrealistic supervised learning rule, on a coordinate transformation task. Mazzoni et al. (1991) used only two output units and a non-binary reward signal directly indicating the amount of error in the outputs and thus accelerating the performance.

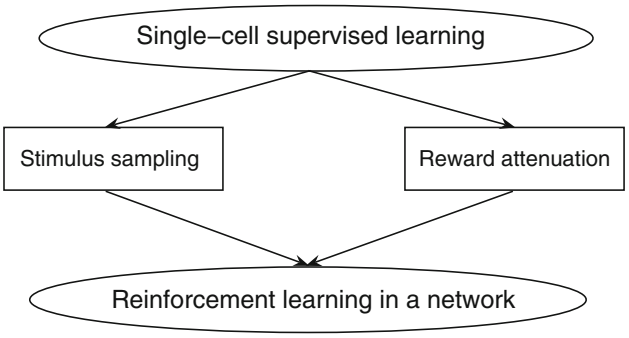

Fig. 5 Suggested modality-independent framework to extend any type of synaptic plasticity suitable for single-neuron supervised learning to reinforcement learning in a network. Stochastic stimulus sampling and reward attenuation are two key components, with the former replacing putative intrinsic exploration mechanisms, and the latter dynamically decreasing the sampling-induced exploration as the performance level increases

Importantly, stimulus sampling also applies to coding schemes in higher cortical areas that are based on population or temporal mean firing rates (Aggelopoulos et al. 2005; Rolls et al. 2006). As noisy synaptic transmission or noisy spike generation would average out for these rate-based codes (unless the noise is correlated) and would thus not be useful for the exploration, stochastic stimulus sampling becomes an even more valuable option that deserves further consideration.

The simplicity of the HRL rule makes our conclusions on stimulus sampling and reward attenuation very general. In particular, they may be applicable to general modality-independent associative learning in the hippocampus (Eichenbaum et al. 1999; Eichenbaum 1999; Buckmaster et al. 2004), as well as to more detailed models of synaptic plasticity. Our construction of a reinforcement learning rule out of classical Hebbian plasticity suggests that a similar framework making universal use of a few basic synaptic plasticity mechanisms could exist in the biological reality. These mechanisms could underly learning within the supervised scenario, where the postsynaptic activity is dominated by some 'teacher input', but also within the reinforcement learning scenario, where only a global reward signal is available. We suggest that whatever the specific biological implementation of a single-cell supervised learning rule might be, it can be combined with stimulus sampling and reward attenuation to generate a corresponding reward-based learning rule for reinforcement learning within a large network (Fig. 5).

4 Methods

4.1 Network architecture

In this first study, we consider feedforward neural networks depicted in Figs. 1 and 4, with n input neurons, m output

neurons, and, for Fig. 4, one, two, or three hidden layers with $n_h = 5$ neurons in each. All neurons are threshold units with binary (0 or 1) output activities y_i for neuron i. In a biological context, the activity of 1 could indicate that the neuron has generated an action potential and has released a neurotransmitter within some integration time period (common to all neurons in the same layer), while 0 would indicate the absence of an action potential within that time. Synaptic efficacies (strengths, weights) J_{ij} from neuron j to neuron i are excitatory, but subject to global inhibition (with separate inhibitory populations in each layer, the results were similar). The synaptic efficacies are continuous and normalized, taking values between 0 and 1. We use $m = 2$ for the simulations of the monkey learning task (Figs. 1, 2), and $m = 1$ otherwise.

Consider a set of P binary stimuli ξ^μ, with components ξ_i^μ=0 or 1 ($i = 1 \ldots n,\ \mu = 1 \ldots P$). The total postsynaptic current to neuron i in the next layer is

$$I_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (J_{ij} - g_I)\xi_j^\mu \tag{1}$$

where n_i is the number of synapses onto neuron i, and g_I defines global inhibition. The neuron is activated ($y_i = 1$) if $I_i > 0$, and inactive ($y_i = 0$) otherwise. The strength of global inhibition was set to $g_I = 0.5$, so that the effective connection strengths range from -0.5 to 0.5. In the presence of multiplicative synaptic bounds, this implies that learning drives the effective synaptic efficacies towards 0, the point of balance between excitation and inhibition (Senn and Fusi 2005).

Replacing ξ_j^μ by y_j, we compute the activities of all cells in the subsequent layers iteratively. Once the states $Z = (z_1, \ldots, z_m)$ of the output neurons have been determined, the same global reward signal $r = r(Z,\ \mu)$ is administered to each synapse. If Z is equal to the desired output configuration for stimulus μ, we set $r = 1$; otherwise, $r = 0$.

4.2 Hebbian reinforcement learning (HRL)

The network learns by changing each synaptic efficacy J_{ij} as a function of the presynaptic activity (x_j), the postsynaptic activity (y_i), and the reward signal $r = r(Z,\ \mu)$. A pure Hebbian learning rule would change the weight proportionally to $(y_i - 0.5)x_i$. Our Hebbian reinforcement learning (HRL) rule changes the weight in the same Hebbian way in the case of reward ($r = 1$) and in an anti-Hebbian way in the case of punishment (i.e., no reward, $r = 0$). Learning in the case of reward is further modulated by a factor of $(1 - r_m)$, where r_m is the running mean of the previously garnered reward (see below). This factor insures that the network learns most from responses that generate the levels of reward farthest from the average. Also, synaptic modifications then properly decrease in response to rewarded stimuli, preventing repeated learning and unlearning of the same stimuli, and learning stops once all the stimuli have produced the desired outputs. Hence, to define the effective weight change we consider the quantity

$$\widetilde{\Delta J_{ij}} = \begin{cases} (1 - r_m)\,\eta\,(y_i - 0.5)\,x_j & \text{if } r = 1 \\ -\eta\,(y_i - 0.5)\,x_j & \text{if } r = 0 \end{cases} \tag{2}$$

where η is some learning rate. In all simulations η was chosen to provide the fastest learning as measured by the median learning time. The specific values used for η are as follows: 0.05 for Fig. 1, 0.05 for HRL and the batch learning and 0.0625 and 0.09 for learning from non-attenuated reward and punishment only, respectively, for Fig. 2. $\eta = 0.0025$ for Fig. 3. For Fig. 4, $\eta = 0.003$ for panels a1 and b1 and $\eta = 0.002$ for panels a2, b2, a3, and b3.

To implement soft synaptic bounds at 0 and 1 during synaptic potentiation and depression, respectively, the effective weight change was

$$\Delta J_{ij} = \begin{cases} \widetilde{\Delta J_{ij}}\,(1 - J_{ij}) & \text{if } \widetilde{\Delta J_{ij}} > 0 \\ \widetilde{\Delta J_{ij}}\,J_{ij} & \text{if } \widetilde{\Delta J_{ij}} < 0 \end{cases} \tag{3}$$

The initial distribution of synaptic efficacies was uniformly random between 0 and 1 except for Figs. 1 and 2, where the network began to learn the set of four novel stimuli with the synaptic weights at the values found during the previous stage of learning the four familiar stimuli. This was intended to reflect the experimental observation (Chen and Wise 1995a; Wirth et al. 2003) that the monkeys would almost always recognize the familiar stimuli correctly when learning the full set of stimuli. The choice of stimulus presented on any trial was (uniformly) random throughout this study, except for the fixed-order batch (Fig. 2a) where all the stimuli were repeatedly presented in the same order.

The average reward r_m was initialized at a uniform random value between 0 and 1 to take into account some reward stochasticity in the network history prior to learning the task at hand. After each stimulus presentation r_m was updated according to

$$\Delta r_m = \lambda(r - r_m) \tag{4}$$

where λ is a small forgetting rate of the previously collected reward within the presentation protocol. We chose a value of λ guaranteeing that more than $3P$ stimulus presentations were required to reach the target average reward value of 0.96, if initially $r_m = 0.5$. When r_m achieves the target value of 0.96, we say that the network has learned and stop the simulation, since with the running mean technique the maximum level of reward equal to 1 cannot be achieved. The value of 0.96 represents less than 5% of incorrect network responses on average. The specific values used for λ are as follows: $\lambda = 0.05$ for learning the four familiar stimuli and $\lambda = 0.07$

for learning the four novel stimuli in Figs. 1 and 2. Note that 0.05 is actually significantly smaller than necessary to simply average over $12 = 3P$ stimulus presentations into the past. This is intended to stabilize the network responses to the familiar stimuli even more than usual to take into account the experimental observation (Chen and Wise 1995a; Wirth et al. 2003) that the monkeys would almost always recognize the familiar stimuli correctly. For Fig. 3, $\lambda = 0.005$, and for Fig. 4, $\lambda = 0.03$.

4.3 Node perturbation (NP)

For a fair comparison, we adapted the original node perturbation learning rule (Jabri and Flower 1992; Cauwenberghs 1993; Werfel et al. 2005) to the case of bounded weights and endowed it with reward saturation as was done for HRL. When the input layer is clamped by a stimulus, the state of the network is computed with synaptic current to neuron i perturbed by a noise signal Δh_i, drawn from a normal distribution with zero mean and variance of σ^2_{NP}. The output y_i of the ith neuron is set to 1 if $I_i + \Delta h_i > 0$ and to 0 otherwise. The noise signals used for each neuron (and on each trial) are independent. After iteratively calculating the neuronal activities, the reward $r = r(Z, \mu)$ is administered and the weights of all the synaptic afferents onto neuron i are updated based on

$$\widetilde{\Delta J_{ij}} = \begin{cases} (1 - r_m)\, \eta_{\mathrm{NP}} \Delta h_i x_j & \text{if } r = 1 \\ -\eta_{\mathrm{NP}} \Delta h_i x_j & \text{if } r = 0 \end{cases} \tag{5}$$

where η_{NP} is a learning rate. To enforce the bounds on the synaptic efficacies, the actual updates ΔJ_{ij} are obtained from (3). Similarly, r_m is updated in the same way as for HRL, with the same parameters. The specific values used for σ_{NP} and η_{NP} are as follows: $\sigma_{\mathrm{NP}} = 0.01$ and $\eta_{\mathrm{NP}} = 1$ for Fig. 1. For Fig. 3, $\sigma_{NP} = 0.0005$ and $\eta_{NP} = 1$. $\sigma_{NP} = 0.0045$ and $\eta_{NP} = 0.3$ for Fig. 4a1, b1; $\sigma_{NP} = 0.002$ and $\eta_{NP} = 0.5$ for Fig. 4a2, b2; $\sigma_{NP} = 0.003$ and $\eta_{NP} = 0.3$ for Fig. 4a3, b3.

4.4 Weight perturbation (WP)

For a fair comparison, we adapted the original weight perturbation learning rule (Widrow and Lehr 1990; Flower and Jabri 1993; Werfel et al. 2005) to the case of bounded weights and endowed it with reward saturation as was done for HRL. Synaptic updates occur twice per learning step: first, a noisy exploratory update, followed by a definitive learning update based on the outcome of the exploration. On the exploratory step current synaptic strengths J_{ij} are updated to become $J_{ij} + \Delta h_{ij}$, and the network output is calculated based on these updated weights. The noise signals Δh_{ij} are drawn from a normal distribution with zero mean and variance of σ^2_{WP}. All Δh_{ij} are independent from neuron to neuron and from trial to trial, but the same sample values are used in the exploratory and learning updates (below).

After obtaining the reward value $r(Z, \mu)$ for the output produced by the exploratory synaptic changes, those changes are undone before the learning update is performed. If $r = 1$, the synapses with presynaptic activity are then changed in the direction of the corresponding provisional noise, whereas the changes are in the opposite direction if $r = 0$. Formally, the original J_{ij} are updated based on

$$\widetilde{\Delta J_{ij}} = \begin{cases} (1 - r_m)\, \eta_{\mathrm{WP}} \Delta h_{ij} x_j & \text{if } r = 1 \\ -\eta_{\mathrm{WP}} \Delta h_{ij} x_j & \text{if } r = 0 \end{cases} \tag{6}$$

where η_{WP} is a learning rate. To enforce the bounds on the synaptic efficacies, the actual updates ΔJ_{ij} are obtained from (3). Similarly, r_m is updated in the same way as for HRL, with the same parameters. The specific values used for σ_{WP} and η_{WP} are as follows: $\sigma_{\mathrm{WP}} = 0.04$ and $\eta_{\mathrm{WP}} = 0.25$ for Fig. 1. $\sigma_{\mathrm{WP}} = 0.003$ and $\eta_{\mathrm{WP}} = 0.5$ for Fig. 4a1, b1; $\sigma_{\mathrm{WP}} = 0.003$ and $\eta_{\mathrm{WP}} = 0.5$ for Fig. 4a2, b2; $\sigma_{\mathrm{WP}} = 0.002$ and $\eta_{\mathrm{WP}} = 0.5$ for Fig. 4a3, b3.

Acknowledgments We thank Dr. S. Fusi for fruitful discussions and his contribution in starting this project. We thank Dr. W. Suzuki for multiple clarifications on the experimental methodology used by Wirth et al. (2003) and valuable comments on the manuscript.

References

Aggelopoulos N, Franco L, Rolls E (2005) Object perception in natural scenes: encoding by inferior temporal cortex simultaneously recorded neurons. J Neurophysiol 93:1342–1357

Asaad W, Rainer G, Miller E (1998) Neural activity in the primate prefrontal cortex during associative learning. Neuron 21:1399–1407

Barto A, Jordan M (1987) Gradient following without back-propagation in layered networks. In: Proceedings of the IEEE first annual conference on neural networks, vol 2. San Diego, pp 629–636

Bayer H, Glimcher P (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47:129–141

Bayley P, Squire L (2002) Medial temporal lobe amnesia: gradual acquisition of factual information by nondeclarative memory. J Neurosci 22:5741–5748

Brasted P, Wise S (2004) Comparison of learning-related neuronal activity in the dorsal premotor cortex and striatum. Eur J Neurosci 19:721–740

Buckmaster C, Eichenbaum H, Amaral D, Suzuki W, Rapp P (2004) Entorhinal cortex lesions disrupt the relational organization of memory in monkeys. J Neurosci 24:9811–9825

Cahusac P, Rolls E, Miyashita Y, Niki H (1993) Modification of the responses of hippocampal neurons in the monkey during the learning of a conditional spatial response task. Hippocampus 3:29–42

Cauwenberghs G (1993) A fast stochastic error-descent algorithm for supervised learning and optimization. In: Giles C, Hanson S, Cowan J (eds) Advances in neural information processing systems, vol 5. Morgan Kaufmann, San Mateo, pp 244–251

Chen L, Wise S (1995a) Neuronal activity in the supplementary eye field during acquisition of conditional oculomotor associations. J Neurophysiol 73:1101–1121

Chen L, Wise S (1995b) Supplementary eye field contrasted with the frontal eye field during acquisition of conditional oculomotor associations. J Neurophysiol 73:1122–1134

Chialvo D, Bak P (1999) Learning from mistakes. Neuroscience 90:1137–1148

Daw N, Doya K (2006) The computational neurobiology of learning and reward. Curr Opin Neurobiol 16:199–204

Doya K (2008) Modulators of decision making. Nat Neurosci 11:410–416

Doya K, Sejnowski T (1998) A computational model of birdsong learning by auditory experience and auditory feedback. In: Brugge J, Poon P (eds) Central auditory processing and neural modeling. Plenum Press, New York, pp 77–88

Eichenbaum H (1999) Cortical-hippocampal networks for declarative memory. Nat Rev Neurosci 1:41–50

Eichenbaum H, Dudchenko P, Wood E, Shapiro M, Tanila H (1999) The hippocampus, memory, and place cells: is it spatial memory or a memory space? Neuron 23:209–226

Fiete I, Seung H (2006) Gradient learning in spiking neural networks by dynamic perturbation of conductances. Phys Rev Lett 97:048104

Flower B, Jabri M (1993) Summed weight neuron perturbation: an $\mathcal{O}(n)$ improvement over weight perturbation. In: Giles C, Hanson S, Cowan J (eds) Advances in neural information processing systems, vol 5. Morgan Kaufmann, San Mateo, pp 212–219

Hebb O (1949) The organization of behavior. Wiley, New York

Hertz J, Krogh A, Palmer R (1991) Introduction to the theory of neural computation. Addison-Wesley, Redwood City

Jabri M, Flower B (1992) Weight perturbation: an optimal architecture and learning technique for analog VLSI feedforward and recurrent multilayered networks. IEEE Trans Neural Netw 3:154–157

Kobayashi Y, Okada K (2007) Reward prediction error computation in the pedunculopontine tegmental nucleus neurons. Ann New York Acad Sci 1104:310–323

Mazzoni P, Andersen R, Jordan M (1991) A more biologically plausible learning rule for neural networks. Proc Natl Acad Sci USA 88:4433–4437

McClure S, Daw N, Montague P (2003) A computational substrate for incentive salience. Trends Neurosci 26:423–428

Mitz A, Godschalk M, Wise S (1991) Learning-dependent neuronal activity in the premotor cortex: activity during the acquisition of conditional motor associations. J Neurosci 11:1855–1872

Montague P, Dayan P, Person C, Sejnowski T (1995) Bee foraging in uncertain environments using predictive hebbian learning. Nature 377:725–728

Montague P, Dayan P, Sejnowski T (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16:1936–1947

Pasupathy A, Miller E (2005) Different time courses of learning-related activity in the prefrontal cortex and striatum. Nature 433:873–876

Rescorla R, Wagner A (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black AH, Prokasy WF (eds) Classical conditioning II: Current research and theory. Appleton-Century-Crofts, New York, pp 64–99

Rolls E, Franco L, Aggelopoulos N, Jerez J (2006) Information in the first spike, the order of spikes, and the number of spikes provided by neurons in the inferior temporal visual cortex. Vis Res 46: 4193–4205

Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 65:386–408

Rumelhart D, Durbin R, Golden R, Chauvin Y (1996) Backpropagation: the basic theory. In: Smolensky P, Mozer M, Rumelhart D (eds) Mathematical perspectives on neural networks. Lawrence Erlbaum Associates, Hillsdale, pp 533–566

Schönberg T, Daw N, Joel D, O'Doherty J (2007) Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. J Neurosci 27:12860–12867

Schultz W (2002) Getting formal with dopamine and reward. Neuron 36:241–263

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

Schultz W, Tremblay L, Hollerman J (2003) Changes in behavior-related neuronal activity in the striatum during learning. Trends Neurosci 26:321–328

Senn W, Fusi S (2005) Convergence of stochastic learning in perceptrons with binary synapses. Phys Rev E Stat Nonlinear Soft Matter Phys 71:061907

Seung H (2003) Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. Neuron 40:1063–1073

Seymour B, O'Doherty J, Dayan P, Koltzenburg M, Jones A, Dolan R, Friston K, Frackowiak R (2004) Temporal difference models describe higher-order learning in humans. Nature 429:664–667

Stark C, Bayley P, Squire L (2002) Recognition memory for single items and for associations is similarly impaired following damage to the hippocampal region. Learn Mem 9:238–242

Stark C, Squire L (2003) Hippocampal damage equally impairs memory for single items and memory for conjunctions. Hippocampus 13:281–292

Sutton R, Barto A (1981) Toward a modern theory of adaptive networks: expectation and prediction. Psychol Rev 88:135–170

Sutton R, Barto A (1998) Reinforcement learning: an introduction. MIT Press, Cambridge

Suzuki W (2007) Integrating associative learning signals across the brain. Hippocampus 17:842–850

Vargha-Khadem F, Gadian D, Watkins K, Connelly A, Van Paesschen W, Mishkin M (1997) Differential effects of early hippocampal pathology on episodic and semantic memory. Science 277:376–380

Vasilaki E, Fusi S, Wang X, Senn W (2009) Learning flexible sensorimotor mappings in a complex network. Biol Cybern 100:147–158

Werfel J, Xie X, Seung H (2005) Learning curves for stochastic gradient descent in linear feedforward networks. Neural Comput 17:2699–2718

Wickens J, Horvitz J, Costa R, Killcross S (2007) Dopaminergic mechanisms in actions and habits. J Neurosci 27:8181–8183

Widrow B, Lehr M (1990) Thirty years of adaptive neural networks: perceptron, madaline, and backpropagation. Proc IEEE 78:1415–1442

Williams R (1992) Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach Learn 8:229–256

Williams Z, Eskandar E (2006) Selective enhancement of associative learning by microstimulation of the anterior caudate. Nat Neurosci 9:562–568

Wirth S, Yanike M, Frank L, Smith A, Brown W, Suzuki W (2003) Single neurons in the monkey hippocampus and learning of new associations. Science 300:1578–1581

Xie X, Seung H (2004) Learning in neural networks by reinforcement of irregular spiking. Phys Rev E Stat Nonlinear Soft Matter Phys 69:041909