J Neurophysiol 125: 2034–2037, 2021.
First published April 28, 2021; doi:10.1152/jn.00712.2020

american physiological society

JNP JOURNAL OF NEUROPHYSIOLOGY

NEURO FORUM

Computational Neuroscience

Classic Hebbian learning endows feed-forward networks with sufficient adaptability in challenging reinforcement learning tasks

Thomas F. Burns

Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa, Japan

Abstract

A common pitfall of current reinforcement learning agents implemented in computational models is in their inadaptability postoptimization. Najarro and Risi [Najarro E, Risi S. *Proc 33rd Conf Neural Inf Process Systems (NeurIPS 2020)*. 2020: 20719–20731, 2020] demonstrate how such adaptability may be salvaged in artificial feed-forward networks by optimizing coefficients of classic Hebbian rules to dynamically control the networks' weights instead of optimizing the weights directly. Although such models fail to capture many important neurophysiological details, allying the fields of neuroscience and artificial intelligence in this way bears many fruits for both fields, especially when computational models engage with topics with a rich history in neuroscience such as Hebbian plasticity.

artificial neural network; computational model; Hebbian learning; learning; reinforcement learning

Development of artificial reinforcement learning agents has grown in popularity and prominence in recent years due to their remarkable near- or better-than-human performance on complex tasks like game-playing. However, such agents—even those which master multiple tasks—are still far less adaptable than humans (1) and can suffer from overspecializing in tasks (2).

Najarro and Risi (3) draw inspiration from neuroscience in a way few in machine learning have previously: instead of optimizing the weights in their artificial neural networks, they optimize the coefficients of Hebbian learning rules which dynamically control the weights at runtime. Their model is still a classic feed-forward policy network (i.e., sensory information is the input and the agent's next action is the output); however, the network never receives any reward signals, and in each task episode, the network's weights start from a random state, even after training. In contrast, similar models explicitly guide the network using a reward function to steer the behavior toward some chosen goal during training, then behavior is embedded in the network by fine-tuning its weights (and thereafter leaving them static) to maximize the attainable reward at runtime. Najarro and Risi's (3) networks instead learn a range of diverse synaptic plasticity rules which—once learnt—allow weights to reliably converge from random values toward reward-attaining values for the task in noisy dynamical contexts.

Although the idea of learning synaptic plasticity rules to optimize artificial neural networks is not new, the optimization approach used by Najarro and Risi (3) is, and their demonstration of better adaptability than current reinforcement learning algorithms without an explicit reward is novel and surprising.

Najarro and Risi (3) begin by declaring a common Hebbian rule which will govern the change in weight between each pair of neurons in their feedforward networks:

$$\Delta w_{ij} = \eta_w(A_w o_i o_j + B_w o_i + C_w o_j + D_w),$$

where o_i and o_j are the pre- and postsynaptic neuron activations, respectively, and the remainder of the coefficients are unique to every pre-post pair and optimized by an evolutionary algorithm. The η_w term is a learning rate parameter, controlling the speed at which the weight will change; $A_w o_i o_j$ is the only purely Hebbian term, following the classic "fire together, wire together" principle; the $B_w o_i$ and $C_w o_j$ terms provide an adaptive, activity-dependent bias for the pre- and postsynaptic neurons, respectively; and D_w is a static bias term (making the weight inhibitory or excitatory). All of the optimized parameters take values in the interval [0,1] and are denoted by the vector h. After the networks' weights are initialized randomly in the interval [−0.1, 0.1], its Hebbian coefficients h (which are independent for each synapse) are set and the network is tested in one of two environments, a

Correspondence: T. F. Burns (t.f.burns@gmail.com).
Submitted 22 December 2020 / Revised 26 April 2021 / Accepted 26 April 2021

2 D car racing environment or a 3 D locomotion environment (see Fig. 1*A*). At the conclusion of a test episode, fitness is measured and the Hebbian coefficients *h* are optimized by an evolutionary algorithm to selectively iterate on the best selections of coefficients.

Before discussing the test environments, a small note: in typical biological intelligent systems, learning is not often optimized for single tasks or environments. Complex behaviors and tasks are often learnt as collections such that intuitions or experience from different behaviors or tasks can be shared and composed for efficiency in cognition and learning (1). Indeed, the term "meta-learning" in the machine learning literature often refers to learning a battery of tasks and autonomously exploiting knowledge of distinct tasks or domains in learning from or reacting to novel tasks.

In Najarro and Risi (3), the two environments tested were a 2D car racing game and a 3D four-legged robot walking task. In both environments, the fitness of the network was measured as how far the car or robot successfully traveled in the environment in a given period of time. After optimizing the Hebbian coefficients via their evolutionary procedure, both tasks could be performed at a comparable level to state-of-the-art reinforcement learning algorithms. However, in the robot environment, they performed two key experiments: *1*) they damaged the left front leg of the robot (during optimization the robot was always whole or only had damage to the right front leg); and *2*) they performed a simulated ablation mid-task by zeroing a portion of the networks' weights. The performance of the network in these two experiments in the 3D robot locomotion environment (leg damage and simulated ablation) is where the authors probe the model's "meta-learning" ability. And although these "meta-learning" abilities may seem unsophisticated or trivial, traditional reinforcement learning models and feed-forward policy networks are almost completely incapable of even basic performance in these situations for two reasons: *1*) they are explicitly optimized using a reward function during training, and therefore are biased toward those rewarded goals instead of learning, for example, more general and/or stable internal representations, dynamics, and control; and *2*) each of the network's weights are uniquely irreplaceable and unrecoverable if and once lost. These kinds of inabilities have previously been identified as overfitting (2) but also come in other forms and are quite fundamental and pervasive (1).

Najarro and Risi (3) reported significant performance improvements on the leg damage and simulated ablation experiments compared to traditional reinforcement-learners. When the left front leg of the simulated 3D robot was damaged (not seen during optimization), the robot could barely move if the network optimized its weights directly in the traditional way (68 ± 56 a.u.), whereas it could move an order of magnitude farther if the network optimized the Hebbian coefficients (452 ± 95 a.u). Although this was still half the distance traveled of the situations the network had been optimized for (showing a limitation in the network's meta-learning ability), it shows how the Hebbian learning rules collectively were able to generalize from no damage and damage to the right front leg in order to walk with damage to the left front leg. This is in stark contrast to the statically fine-tuned weights, which showed essentially no generalization of learning and probably overfit the specific tasks presented during optimization (walking without damage and with damage only to the right front leg). In fact, given the geometric symmetry of the robot's legs, it is quite startling how badly the traditional approach performed given the damage was simply reflected along one morphological axis.

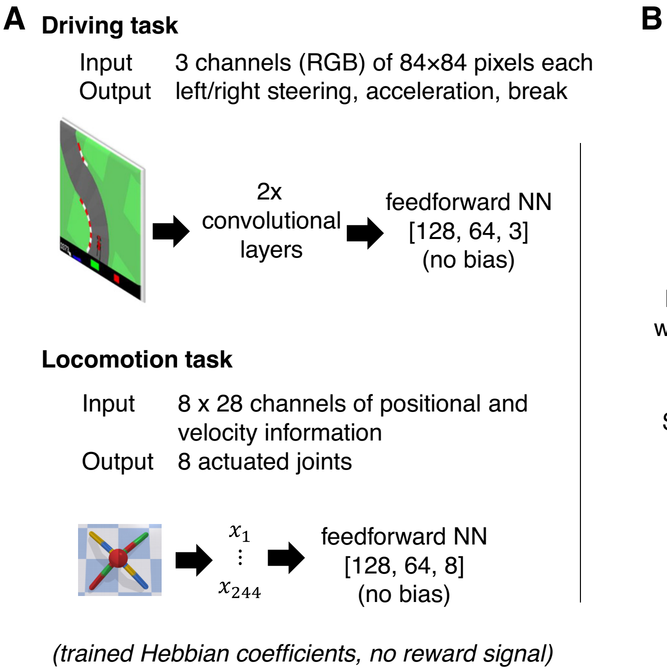
A
Driving task
Input 3 channels (RGB) of 84×84 pixels each
Output left/right steering, acceleration, break
2x convolutional layers
feedforward NN [128, 64, 3] (no bias)
Locomotion task
Input 8 x 28 channels of positional and velocity information
Output 8 actuated joints
x_1
⋮
x_{244}
feedforward NN [128, 64, 8] (no bias)
(trained Hebbian coefficients, no reward signal)
B
Structured, biologically-realistic sensory input
Networks with:
Hebbian learning with reward signals
Balanced structural and adaptive components
Spiking (using the time domain)
Genetic bottlenecks
Dale's Law and output positivity
Dendrites

Figure 1. *A*: schematic overview of Najarro and Risi's (3) tasks and network structure. *B*: summary of proposed extensions and additions for future models. NN, neural network; RGB, red green blue.

Learning the Hebbian learning rules also enabled the network to quickly recover from a partial zeroing of its weights mid-task—after momentarily faltering, the robot continued walking and the weights were functionally restored. This suggests the Hebbian learning rules were not only able to collectively learn and generalize the walking task but also learn and maintain internal network dynamics critical to the task. Although Najarro and Risi (3) didn't directly test the performance of the traditional reinforcement-learners with the same partial weights zeroing, it is reasonable to expect the control and dynamics to be completely lost, and this would be irrecoverable since the weights are static and not dynamic.

Remarkably, these networks also gained rewards in traditional reinforcement learning environments without an explicit reward signal. I speculate it may be exactly the lack of an explicit reward signal which supplies these meta-learning abilities. Without this signal, the network is free to explore the learning rule space (of all synapses independently) in any nonlinear fashion it likes through the evolutionary procedure (not necessarily in a smooth, gradient-based, or differentiable manner). Then, the training selects for stable, nonspecific success (e.g., in driving or locomotion distance), i.e., there is no specified reward or goal and the network has to perform stably in the presence of random dynamic perturbations—procedurally generated, random racing tracks with unexpected turns in the case of the driving task, and changing, partially damaged robot morphologies in the case of the locomotion task—as well as randomly initialized weights. I, therefore, hypothesize these training procedures allow meta-Hebbian learners to develop more generalized and robust (if still rudimentary) internal models of how their action choices affect the environment-agent feedback loop compared to reinforcement learners. I speculate this feat is by virtue of their broader, less constrained exploration of the environment and their training requirement to self-generate internal dynamic stability in their weights. Meta-Hebbian learners then exploit their rudimentary (but far superior) internal model to improve their genetic fitness, avoiding overspecialization in (rewarded) tasks and improving stable generalization. This hypothesis, if correct, would also explain why the meta-Hebbian learner walked approximately half as far as the traditional reinforcement learner in Najarro and Risi's (3) robot locomotion task—the meta-Hebbian learner was carrying the additional cognitive burden of a rudimentary internal model whereas the reinforcement learner could "put their blinders on" and focus purely on the task at hand.

Nonetheless, biological neural networks clearly do utilize rewards during their learning and development, even if some of their early-life cognitive abilities have strong genetic bases (4). One immediate question, therefore, is whether reward-based extensions of Hebbian learning (5, 6) which are not blind to the goal of tasks perform even better in Najarro and Risi's (3) framework.

Another area for improvement is in the evolutionary procedure used for training. Although newborn animals do express complex behaviors without much or any experience of the world (e.g., breathing or walking), it is unlikely that between each pair of neurons a unique synaptic learning rule is genetically coded, as in the Najarro and Risi (3) model, and that these learning rules are solely responsible for cognitive abilities. It is much more likely learning rules are partly shared between synapses and, along with network structure, are compressed via a "genomic bottleneck" (4)—and then improved further postbirth via learning and development. Indeed, we know there exists a very high degree of structure in, for example, sensory systems, all the way from the level of the external sensor to the innermost cortical representations and circuits, consisting of a range of structural rules with neuroanatomical features and diverse sets of interneurons playing different dynamical roles in the network (7). Such structure is completely absent in the Najarro and Risi (3) model, since they explicitly aimed to study the effects of random networks. However, what may make their framework more interesting from a neuroscientific perspective is by inserting more structure into the network and studying the trade-offs between structure and adaptive mechanisms. Another approach—which the original authors seem to already be pursuing (8)—could be to test different types of "genomic bottlenecks," perhaps encoding varying balances of structure (e.g., network topology) and adaptive mechanisms (e.g., synaptic learning rules). In this context, an interesting question to ask might be: if there is a set of genetically encoded Hebbian learning rule parameters which encode reward-attaining behaviors, why not encode these behaviours in weights or other static features directly instead of a learning rule or other adaptive mechanisms? Although I hypothesized earlier about the possible generalization advantage of meta-Hebbian learners, it is also possible some trade-offs between generalization and efficiency are beneficial to learning and genetic fitness. These trade-offs might also be affected by factors like the life span of the agent, its reproductive cycle, and its early stages of development.

From the neurophysiology perspective, there are still more overlooked details. I will focus on only three more which I believe—if worked on collaboratively by neuroscientists and machine learning researchers—could provide value to both fields (summarized in Fig. 1*B*).

First, neuron and network structure: Najarro and Risi (3) implicitly model dendriteless neurons with firing rates in feedforward, all-to-all connected layers using synchronous computation. Although this is by far the most common modeling paradigm in modern deep learning, the absence of spikes and dendrites, the breaking of Dale's Law and output (firing rate) positivity, and the artificial, arbitrary, "tabula rasa"-like network topology leaves wide scope for experimentation. I suggest, for example, translating Najarro and Risi's (3) work into a spiking neural network model with dendrites. Spikes endow great computational power via utilization of the time domain (9); temporal coding might afford greater efficiency. Dendrites provide many essential and primary computations in biological networks (10), thereby increasing the computational capacity of individual neurons and the network as a whole. Incorporating these and other biological features (e.g., reward-based extensions of Hebbian learning) may improve model performance while simultaneously elucidating computational characteristics of the underlying biology.

Relatedly, applying neuron-to-neuron Hebbian learning rules to neurons modeled using firing rates is slightly dubious; some might rather interpret these as small populations of neurons, and thus it might be inappropriate to apply a learning rule mostly studied between discrete biological neuron pairs. Experimentalists could further develop the technique and analysis of cultured multilayer networks (11) to verify how similar population-to-population Hebbian mechanisms are in feedforward network models.

Second is Najarro and Risi's (3) use of the hyperbolic tangent as their neurons' activation function. This choice allowed their Hebbian parameters $h \in [0, 1]$ to effectively take on both negative and positive values, in an adaptive way, depending on the activation value (in the interval of $[-1, 1]$). This means their neurons could flip sign activation-to-activation, sometimes representing an excitatory neuron (population) and sometimes representing an inhibitory neuron (population). This is very unnatural. Arguably worse, negative output firing rates were implicitly permitted in the model. If future computational work aims to improve biological realism, modelers should attempt to constrain both the flipping and output negativity.

Third, Najarro and Risi (3) show their algorithm optimizes networks to have continuous, Gaussian distributions of Hebbian coefficients. Although there is certainly a heterogeneity of plasticity and learning rules in the brain, it is unclear whether these rules form continuous distributions. And, where continuous distributions do exist in the brain, such as in the resulting synaptic weight distributions, the distributions are far from universally Gaussian (12). Indeed, there may be strong computational implications of this—in a related work, Palm et al. (8) hypothesize limiting such distributions through a "genomic bottleneck" may help the network generalize its learning between tasks. Further adding specific nonlinear terms, for example, with lognormal weight transformations, may reveal added performance advantages. Further work could also be done by experimentalists to determine how continuous or discrete the distribution of certain plasticity and learning rules actually are to provide stricter bounds for such models.

There are still many more biological details which this and many models ignore, and many of these details are important for function. The three details mentioned above (neuron and network structure, dynamic and intrinsic output values of neurons, and distributions of plasticity and learning rules) represent just a few which I personally find most fundamental or interesting. Combined with further improvements in the evolutionary or other procedures used for training, and the potential extension to reward-based Hebbian learning, I believe there are clear opportunities for neuroscientists and machine learning researchers to collaborate and jointly discover knowledge useful to both fields.

In one final suggestion for collaboration between these fields, I would point out to neuroscientists how capable and performant these models are, and how totally transparent and manipulable they are—they can and should be used more frequently as "model organisms." And to machine learning researchers, I would point out how making a single change to a network's dynamics and optimization (in a more biologically plausible direction) gave considerable advantages over state-of-the art methods—neuroscience and biology have plenty more to offer in terms of adaptability and performance.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

T.F.B. conceived and designed research; T.F.B. drafted manuscript; T.F.B. edited and revised manuscript; T.F.B. approved final version of manuscript.

REFERENCES

1. **Lake BM**, **Ullman TD**, **Tenenbaum JB**, **Gershman SJ.** Building machines that learn and think like people. *Behav Brain Sci* 40: e253, 2017. doi:10.1017/S0140525X16001837.
2. **Zhang C**, **Vinyals O**, **Munos R**, **Bengio S.** A study on overfitting in deep reinforcement learning. *arXiv*, 1804.068932018, 2018.
3. **Najarro E**, **Risi S.** Meta-learning through Hebbian plasticity in random networks. *Proc 33rd Conf Neural Inf Process Systems (NeurIPS 2020)*. 2020: 20719–20731, 2020. https://arxiv.org/abs/2007.02686.
4. **Zador AM.** A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat Commun* 10: 3770, 2019. doi:10.1038/s41467-019-11786-6.
5. **Frémaux N**, **Gerstner W.** Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Front Neural Circuits* 9: 85, 2016. doi:10.3389/fncir.2015.00085.
6. **Miconi T.** Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *eLife* 6: e20899, 2017. doi:10.7554/eLife.20899.
7. **Burns TF**, **Rajan R.** Sensing and processing whisker deflections in rodents. *PeerJ* 9: e10730, 2021. doi:10.7717/peerj.10730.
8. **Palm RB**, **Najarro E**, **Risi S.** Testing the genomic bottleneck hypothesis in Hebbian meta-learning (Preprint). *arXiv*, 2011.06811, 2011.
9. **Maass W.** Lower bounds for the computational power of networks of spiking neurons. *Neural Comput* 8: 1–40, 1996. doi:10.1162/neco.1996.8.1.1.
10. **London M**, **Hausser M.** Dendritic computation. *Annu Rev Neurosci* 28: 503–532, 2005. doi:10.1146/annurev.neuro.28.061604.135703.
11. **Barral J**, **Wang X-J**, **Reyes AD.** Propagation of temporal and rate signals in cultured multilayer networks. *Nat Commun* 10: 3969, 2019. doi:10.1038/s41467-019-11851-0.
12. **Buzsáki G**, **Mizuseki K.** The log-dynamic brain: how skewed distributions affect network operations. *Nat Rev Neurosci* 15: 264–278, 2014. doi:10.1038/nrn3687.