Backpropagation and the Brain

Giuseppe di Pellegrino
Department of Psychology, University of Bologna
Cognition and Neuroscience, MSc in Artificial Intelligence (2024/25)

1. Introduction: Learning in Brains and Machines

Humans and animals possess an extraordinary capacity for adaptive learning, enabling them to thrive in complex and dynamic environments. While evolutionary priors such as innate neural architecture provide a foundational basis for behavior, the acquisition of sophisticated skills and knowledge often requires extensive learning. This learning is implemented through experience-dependent modifications in the functional mapping between neural inputs and outputs.

At the neuronal level, long-term changes in synaptic strength are widely accepted as the biological substrates of learning. Understanding how such plasticity scales to behaviorally relevant outcomes requires bridging insights from biological neural circuits and computational models of artificial learning.

2. Synaptic Plasticity: Biological Underpinnings of Learning

Synaptic plasticity, the modification of synaptic weights based on activity, is pervasive across the brain. Early theories by Donald Hebb (1949) posited that connections between co-activating neurons are reinforced. This Hebbian principle underlies many computational models of learning.

Empirical studies have demonstrated synaptic plasticity in the hippocampus, notably in the form of long-term potentiation (LTP), first reported by Bliss and Lomo (1973). LTP is input-specific and associative, relying on both presynaptic activity (e.g., glutamate release) and postsynaptic depolarization, mediated through NMDA receptor activity.

These mechanisms are present in diverse brain regions, including the amygdala during fear conditioning, and the cerebellum in supervised motor learning. The cerebellum exemplifies supervised learning by comparing predicted and actual sensory outcomes to update synaptic weights, guided by climbing-fiber error signals.

3. The Challenge of Network-Level Learning

While synaptic plasticity explains localized synaptic changes, it does not fully account for coordinated network-level learning. Effective learning systems require mechanisms for credit assignment—determining how synaptic changes in upstream layers contribute to overall behavioral success.

Machine learning addresses this problem through algorithmic coordination of synaptic updates to minimize a predefined error function. The backpropagation of error (Rumelhart et al., 1986) is the most successful of such algorithms in deep neural networks, systematically distributing error information through recursive computations across layers.

4. Backpropagation in Artificial Neural Networks

Backpropagation leverages exact causal relationships between weights and outputs to calculate precise gradients. It computes weight updates using the chain rule of calculus, enabling efficient, layer-by-layer adjustment of synaptic weights based on the network's output error.
It supports a range of learning paradigms:

Supervised learning: error = target - prediction

Reinforcement learning: temporal difference error

Unsupervised learning: reconstruction or prediction error

Key features relevant to biological plausibility include:

Synapse-specific learning, echoing STDP

Requirement for feedback connections to transmit error signals

However, these feedback pathways in biological systems may be delayed, coarse, or scalar, contrasting with the precise gradient information required by backpropagation.

5. Biological Limitations of Backpropagation

Several fundamental challenges undermine the biological plausibility of backpropagation:

Supervision Signals: The brain lacks explicit, centralized supervision. Instead, it likely relies on a combination of weak, distributed signals (e.g., reward, context, prediction).

Neural Signal Representation: Artificial neurons output continuous values; biological neurons transmit information via discrete spikes, making derivative-based learning non-trivial.

Signed Error Signals: Backprop relies on error signals of variable magnitude and sign, which are difficult to implement biologically in a deep network.

Weight Symmetry: The algorithm requires symmetric forward and backward weights, a constraint biologically implausible due to the weight transport problem.

Locality of Learning: Biological synapses only access local signals, whereas backpropagation demands global error computations that are inaccessible to individual synapses.

Role of Feedback: In the brain, feedback is not limited to training but also modulates activity during inference and cognition, unlike in most artificial networks.

6. Biologically Inspired Alternatives to Backpropagation

Although there is no direct evidence that the brain employs backpropagation-like algorithms, artificial neural networks trained with backpropagation have successfully replicated neural activity patterns observed in regions such as the posterior parietal cortex, primary motor cortex, and visual cortex. These findings imply that, despite biological constraints, backprop-like learning mechanisms remain plausible from a representational standpoint.

Recent efforts aim to approximate backpropagation in biologically plausible ways. One such approach is Feedback Alignment (FA), where random fixed feedback weights replace symmetric connections. Despite the noise, feedforward weights align with these feedback pathways over time, enabling learning.

Another promising framework is the NGRAD (Neural Gradient Representation by Activity Differences) hypothesis.

The NGRAD hypothesis offers a biologically plausible alternative to backpropagation by proposing a mechanism that relies solely on local activity differences rather than the explicit propagation of error derivatives.
In conventional backpropagation, two distinct signals must be transmitted through the network: neuronal activations (forward pass) and error gradients (backward pass). This dual-signal requirement presents major biological implausibilities, such as the need for synaptic weight symmetry, and non-local credit assignment.

NGRAD circumvents these challenges by postulating that feedback from higher-order cortical areas—which may represent task goals, contextual information, or supervisory signals—can directly modulate the activity of neurons in lower-order areas. Crucially, this modulation leads to a difference in activity states: one representing the neuron's feedforward response to the input, and the other representing its feedback-influenced activity.

This difference in activity functions as a proxy for error. Rather than computing gradients explicitly, the synapse observes how its associated neuron's activity changed as a result of feedback. This local, temporally distinct signal (before vs. after feedback) serves as a basis for synaptic weight updates, much like how a gradient would guide learning in artificial networks.

In this way, NGRAD preserves the computational benefits of backpropagation—specifically, efficient credit assignment across layers—while adhering to the constraints of biological plausibility. It relies only on local signals and does not require symmetric connections or global knowledge of the network's error.

In practice, architectures such as multilayer autoencoders serve as testbeds for NGRAD. These models use reconstruction targets and activity differences at each layer to guide learning locally, a method known as Difference Target Propagation (DTP).

7. Conclusions and Future Directions

While backpropagation is a powerful and widely used algorithm in machine learning, its direct implementation in biological systems is constrained by numerous anatomical and physiological limitations. Nonetheless, its conceptual framework continues to inform models of cortical learning.

Emerging algorithms such as NGRAD and DTP offer compelling alternatives that retain core principles of backpropagation, such as efficient credit assignment, while aligning more closely with known biological mechanisms.

Understanding how the brain approximates gradient-based learning remains a fundamental challenge in neuroscience and artificial intelligence. Future research should aim to further elucidate how local activity-driven learning rules can scale to complex, multilayered behaviors in both natural and artificial systems.

Essential references

Lillicrap, T.P., Santoro, A., Marris, L., Akerman, C.J., & Hinton, G. Backpropagation and the brain. Nat. Rev. Neurosci., 2020, 21: 335-346.

Magee, J.C., & Grienberger, C . Synaptic plasticity forms and functions. Ann. Rev. Neurosci., 2020, 43:95–117

Rumelhart, D.E., Hinton, G.E., & Williams, R.J., Learning representations by back-propagating errors. Nature, 1986, 323: 533-536.

Whittington, J.C.R, & Bogacz, R,. Theories of error back-propagation in the brain. Trends Cog. Sci., 2019, 23:235-249.