An energy-based model of folded autoencoders for unsupervised learning in cortical hierarchies

1Heidelberg University, Kirchhoff-Institute for Physics, 2University of Bern, Department of Physiology, 3Institute of Neuroinformatics, UZH / ETH Zurich.

Motivation

Whether the brain uses an optimization scheme like backprop to guide synaptic plasticity in deep hierarchical cortical areas is still an open question. Currently work in progress!

Recently, several models explaining how backprop might be realized in the cortex have been proposed, using predictive coding, inhibitory microcircuitry as well as energy-based and Lagrangian neurodynamics. Here, we extend these models to unsupervised learning and bidirectional (supervised and unsupervised) learning while maintaining a high degree of biological plausibility.

1. Models of error backpropagation

1. Predictive coding

• invert architecture and add input \(\rightarrow\) supervised learning, approximately backprop
• needs phases for learning, plasticity is only active when the network is stationary
• high-level, no implementation details

2. Dendritic microcircuit

• apical compartments encode prediction error
• errors calculated via inhibitory microcircuit
• microcircuit weights trainable to cancel top-layer feedback, no weight transport
• neurons hold both forward and error information
• requires phases for learning

3. Neuronal Least Action

• neuronal dynamics derived from Euler-Lagrange equations + prospective coding
• derived neurodynamics: leaky integrators with look-ahead dynamics \(\phi(u, r) = r + \tau \phi(u + \tau)\) (A)
• no phases, time-continuous backprop (B)
• same error interpretation as the dendritic microcircuit model
• see also: Poster T19 by Kungl, Alos F. et al.

2. Folded autoencoder structure

\[
W: \text{discriminative / forward weights} \\
G_1: \text{generative / backward weights}
\]

Learning useful latent representations through bottleneck, similar to autoencoders.

3. Model and neurophysiological interpretation

Network structure and dynamics condensed in energy function (squared errors):

\[
y \cdot E = \frac{\lambda}{2} \sum_i \| u_i - W_i r_{i-1} \|^2 + \frac{1}{2} \sum_i \| u_i - G_i f_{i+1} \|^2 + \beta C
\]

\(\lambda\): gating of forward and backward flow, \(C\): reconstr. error \(C = \| u_0 - u_0^{\text{target}} \|^2\)

\(g_1\): leak conductance, \(y = 1 + \frac{2}{\beta}\)

Neurosynaptic dynamics derived as gradient descent:

\[
\Delta u_i = -\frac{\partial E}{\partial u_i}, \quad \Delta W_i = -\frac{\partial E}{\partial W_i}, \quad \Delta G_i = -\frac{\partial E}{\partial G_i}
\]

\(u_i\) = \(g_i(W_{i-1} - u_i) + \gamma e_i - \alpha i\)

\(g_i\) = \(g_i(f_i(W_{i+1} - u_i) - \alpha i)\)

\(\gamma\) = \(\gamma f_i(W_{i+1} - u_i)\)

\(e_i\) = \(\frac{\partial E}{\partial e_i}\)

\(\alpha\) = \(\alpha f_i(W_{i+1} - u_i)\)

Both plasticity rules can be interpreted as Urbanczik-Senn type rules:

\(\Delta G_i \propto \gamma f_i(W_{i+1} - u_i)\)

\(\Delta G_i \propto \gamma f_i(W_{i+1} - u_i)\)

\(\Delta G_i \propto \gamma f_i(W_{i+1} - u_i)\)

\(\Delta G_i \propto \gamma f_i(W_{i+1} - u_i)\)

For edge cases (\(\lambda = 0\) or \(\lambda = 1\)), the plasticity approximates error backpropagation, e.g., for \(\lambda = 0\) we get \(\Delta G_i \propto \gamma f_i(W_{i+1} - u_i)\) and \(\Delta W_i \propto \gamma f_i(W_{i+1} - u_i)\)

4. One learning rule, two optimizations

Using the solution of stationary neurodynamics as well as choosing \(\lambda << 1, \lambda > 0\) the plasticity rules can be rewritten as:

\[
\Delta G_i \propto (-1 - \lambda^2)\gamma + \beta C
\]

\(\Delta G_i \propto (-1 - \lambda^2)\gamma + \beta C\)

\(\Delta W_i \propto (-1 - \lambda^2)\gamma + \beta C\)

Test by encoding with discriminative and decoding with generative path:

5. Outlook

• Bidirectional learning by adding cost function in latent layer \(\rightarrow\) currently work in progress!
• Spiking neuron models?

Initial results for classification of MNIST images with stochastic binary neurons and refractory period of 3ms.

Email contact: dodo@kip.uni-heidelberg.de