How and When Random Feedback Works:
A Case Study of Low-Rank Matrix Factorization

Shivam Garg

Stanford University

Santosh S. Vempala

Georgia Tech

Hierarchical Information Processing in the Brain

- Multiple layers of neurons from perception to cognition.
- Learning based on updating synaptic weights.
- Error information may only be available in the later layers.
- How to update earlier layer weights, despite locality constraints?

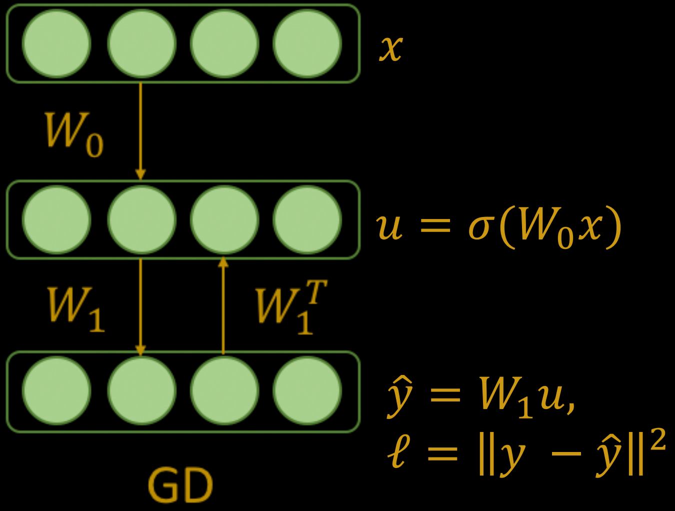
Image source: https://en.wikipedia.org/wiki/Visual_cortex#/media/File:Ventral-dorsal_streams.svg

Gradient descent: Weight Transport Problem

- Artificial NN solution: Use gradient descent via backpropagation.
- GD update:

$$\Delta_{W_0} = \left(\left(W_1^T\,(y - \hat{y})\right) \odot \sigma'(u)\right) x^T$$

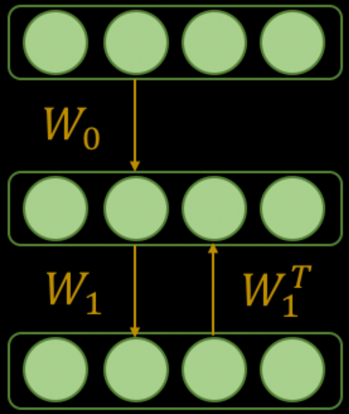
Update to W_0 involves transpose of W_1



Gradient descent: Weight Transport Problem

$$\Delta_{W_0} = \left(\left(W_1^T\left(y - \hat{y}\right)\right) \odot \sigma'(u)\right) x^T$$

- Earlier layer weight updates depend on the transpose of later layer weights.
- Biologically difficult: Requires equi-weighted bidirectional links b/w neurons.
- Weight Transport Problem (Grossberg, 1987).

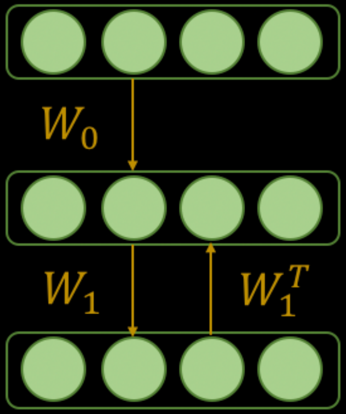


GD

Random Feedback Works

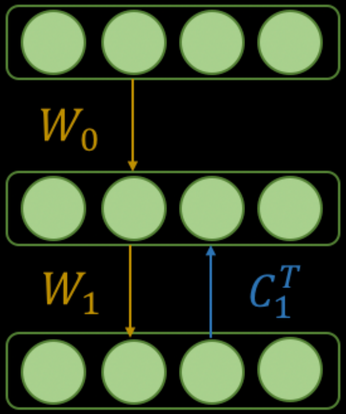
- Replacing weight transpose terms by fixed random matrices also work!
- Feedback Alignment (Lillicrap et. al., 2016).

$$\Delta_{W_0} = \left(\left(W_1^T \, (y \; - \hat{y}) \right) \odot \sigma'(u) \right) x^T$$



GD

$$\Delta_{W_0} = \left(\left(C_1^T (y \; - \hat{y}) \right) \odot \sigma'(u) \right) x^T$$



FA

How and When Random Feedback Works

- Low-rank Matrix Factorization:

Given a matrix $Y_{n \times n}$, find a low-rank factorization $Z_{n \times r}$ $W_{r \times n}$ minimizing the error $\|Y - ZW\|_F^2$.

- Results also hold for training two-layer neural networks with linear activation, assuming isotropic input and target output is a linear function of input.

Summary of Results

- Give convergence analysis shedding light on the alignment dynamics and how they facilitate convergence (for a variant of feedback alignment algorithm).
- FA suboptimal in certain regimes, giving the first provable separation result between FA and GD.
- Representations (matrix Z) found by FA and GD can be almost orthogonal even when their errors $||Y - ZW||_F^2$ are approximately equal.