A biologically-plausible learning rule using reciprocal feedback connections

Mia Cameron[1, 2], Yusi Chen[1,3], Margot Wagner[1] and Terrence J. Sejnowski[1]

[1]Computational Neurobiology Laboratory, Salk Institute for Biological Studies, [2] Department of Mathematics, University of California San Diego
[3]Computational Neuroscience Center, University of Washington, Seattle *:Correspondence: mcameron@ucsd.edu (M.C.)

Locally learning pseudoinverse feedback connections

- Linearized modification of the bio-plausible Recirculation algorithm for autoencoders is equivalent to learning a pair of pseudoinverse weight matrices. [1]
- Can be implemented with random, mean-zero noise at each layer.

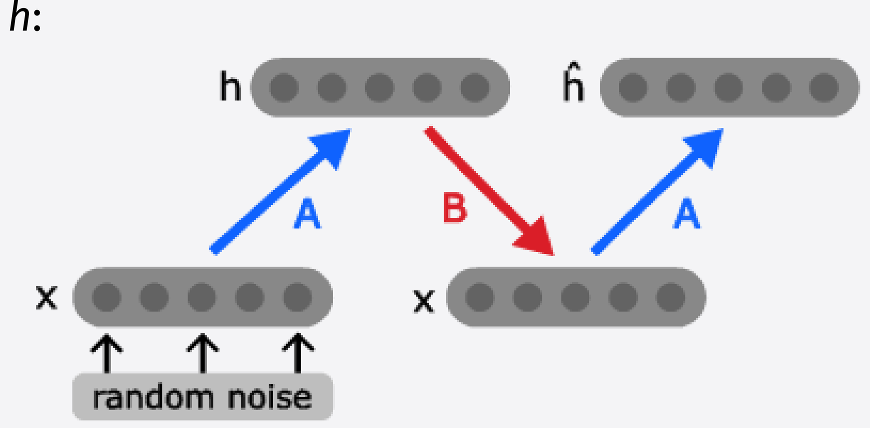
Dynamics:

$h = Ax$	initial hidden representation
$\hat{x} = Bh$	input reconstruction
$\hat{h} = A\hat{x}$	reconstructed hidden representation

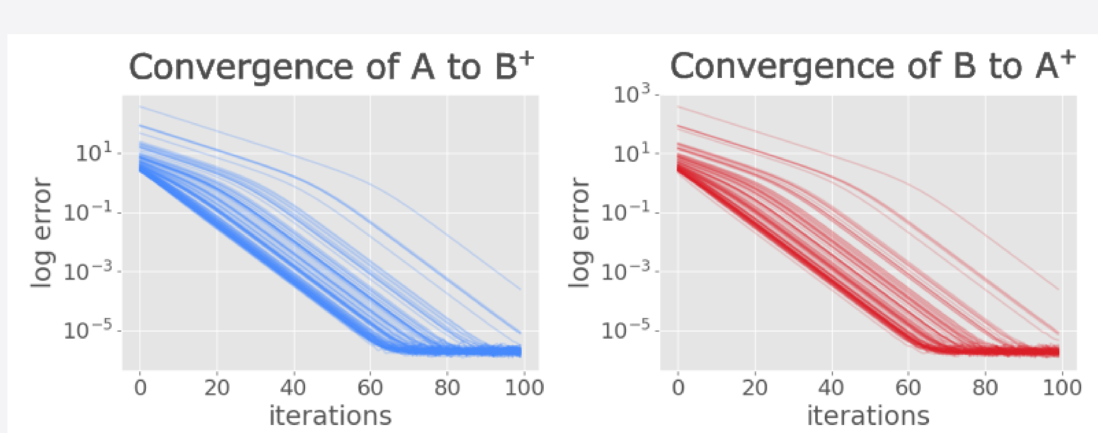
Learning rules:

$\Delta A = (h - \hat{h})x^T$	hidden reconstruction error
$\Delta B = (x - \hat{x})h^T$	input reconstruction error

Unrolling dynamics in time, with original input as x, initial hidden representation as h, reconstructed input as $\hat{x}$ and reconstructed hidden representation as $\hat{h}$:



Simulations of these learning rules (trained concurrently) show convergence to the pseudoinverse:



Def. Moore-Penrose Pseudoinverse

The unique Moore-Penrose pseudoinverse of the $n \times m$ matrix A is the $m \times n$ matrix B satisfying conditions:

1. $ABA = A$ 2. $BAB = B$ 3. $(AB)^T = AB$ 4. $(BA)^T = BA$

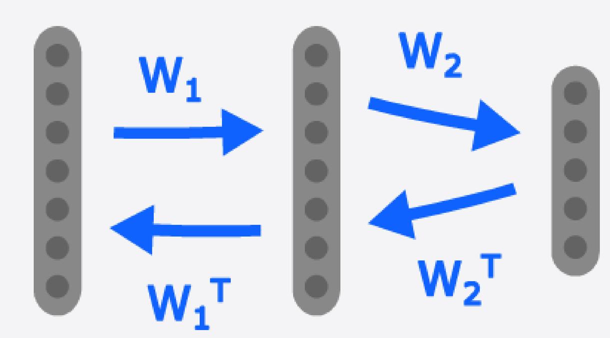
Acknowledgements

We would like to thank Jorge Aldana for assistance with computing resources, and everyone at CNL for helpful comments. This work was supported by the Office of Naval Research Grant N00014-23-1-2069.

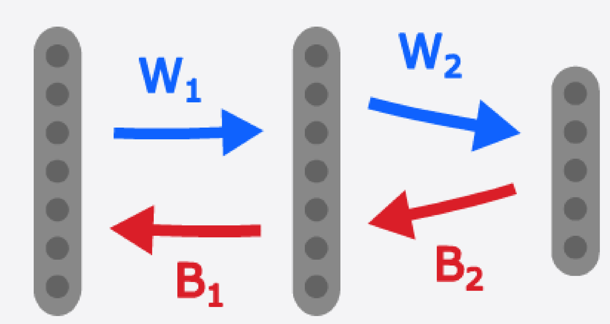
1. Hinton E., G. et al. *AIP* (1988). 2. Crick, F. *Nature* **337**, 129–132 (Jan. 1989). 3. Lillicrap, T. P. et al. *Nat. Rev. Neuros.* **21**, 335–346 (Apr. 2020). 4. Lillicrap, T. P. et al. *Nat. Comm.* **7** (Nov. 2016). 5. Levin, Y. et al. *Nonlinear Analysis: Theory, Methods and Applications* **47**, 1961–1971 (Aug. 2001).

Background

- Problem: Backpropagation is biologically implausible [2]
- Idea: Layer-wise feedback connections [3]



$$\frac{dL}{dW_2} = \frac{dL}{dh_2}\frac{dh_2}{dW_2} = eh_1^T$$



$$\frac{dL}{dW_1} = \frac{dL}{dh_1}\frac{dh_1}{dW_1} = W_2^T e h_0^T$$

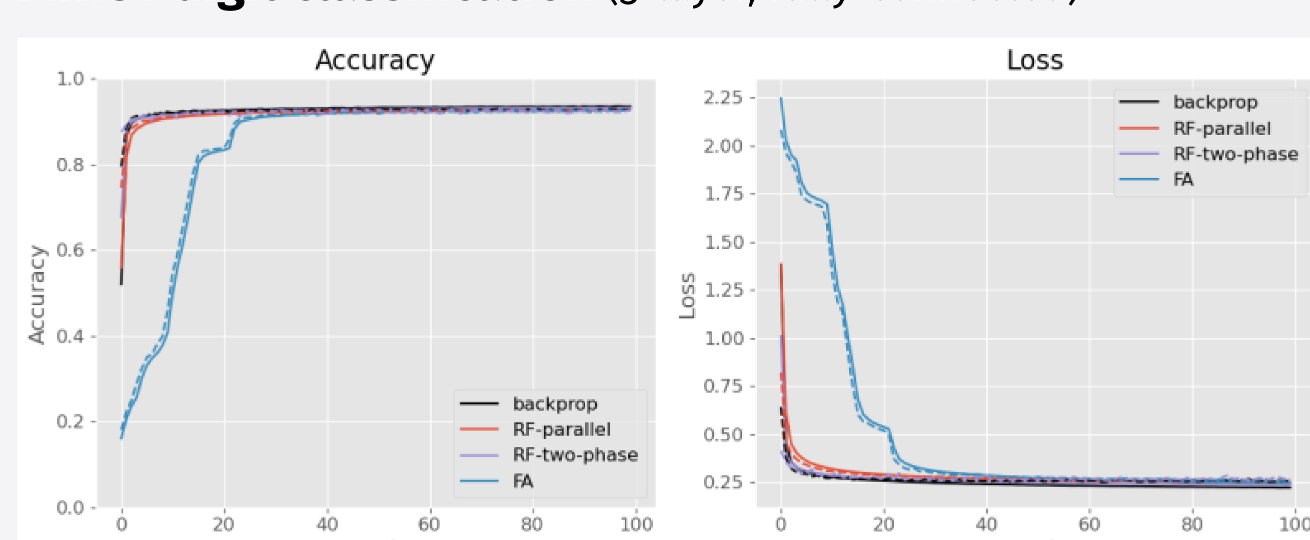
Gradient descent: Separate feedback:

$$\delta_{W_l} = -(J^L_{W_l})^T e h^T_{l-1} \qquad \delta_{W_l} = -\left(\prod_{i=l+1}^{L} B_i\right) e h_{l-1}{}^T$$

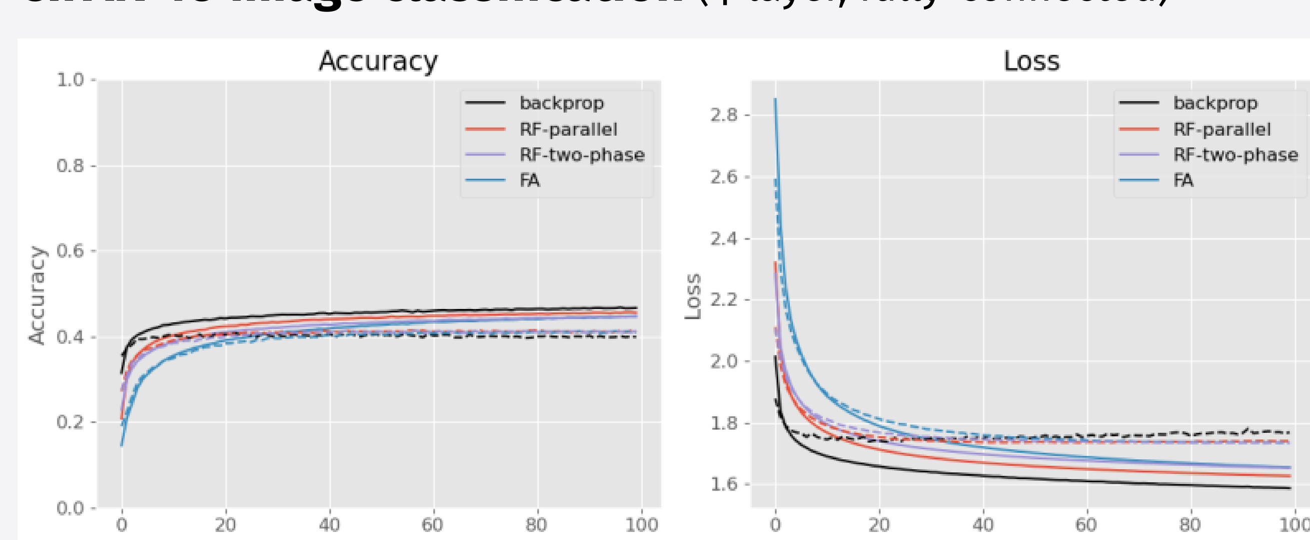
- What choice of B is biologically realistic, and capable of minimizing a global error signal?

Results

MNIST digit classification (5-layer, fully-connected)

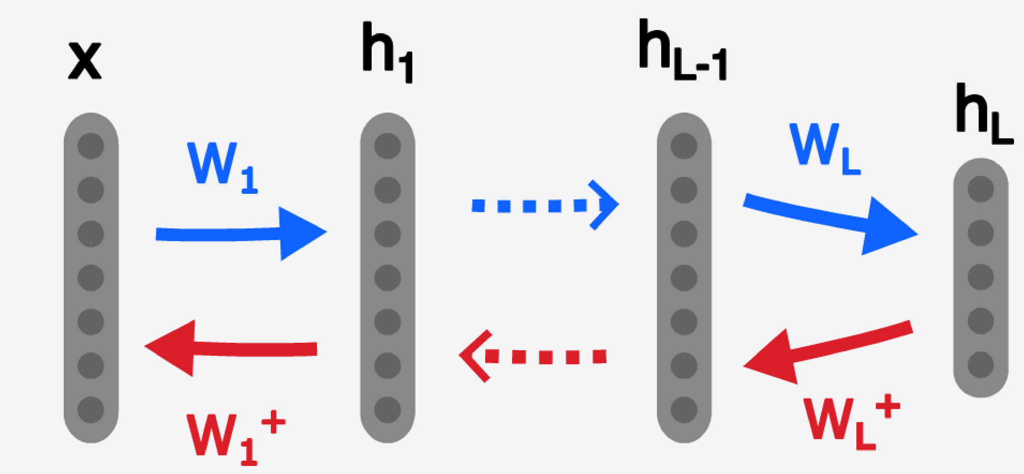


CIFAR-10 image classification (4-layer, fully-connected)



In both CIFAR-10 and MNIST image classification tasks, our method (RF) reaches a similar asymptotic error to backpropagation, with faster rate of convergence than the Feedback Alignment algorithm [4]

Global error minimization using pseudoinverse feedback



Global loss (for one input): $\mathcal{L} = \frac{1}{2}||h_L - h_L^*||^2$

Residual vector: $e = h_L - h_L^*$

Layer-wise Jacobians:

$$J^{\mathcal{L}}_{W_l} = J^e_{W_l} e$$

$$J^e_{W_l} = (h_{l-1} \otimes J^e_{h_l}) \quad \text{w.r.t layer weight matrix}$$

$$J^e_{h_l} = J^e_{h_{l+1}} W_{l+1} = W_L W_{L-1} \ldots W_{l+1} \quad \text{w.r.t layer activation vector}$$

A generalized left inverse can also be defined recursively, corresponding, physically, to the "backwards" application of each layer-wise pseudoinverse to a top-level vector.

Layer-wise left reciprocals:

$$B_{W_l} = (h^+_{l-1} \otimes B_{h_l}) \quad \text{w.r.t layer weight matrix}$$

$$B_{h_l} = W^+_{l+1} B_{h_{l+1}} = W^+_{l+1} \ldots W^+_{L-1} W^+_L \quad \text{w.r.t layer activation vector}$$

Note: $(W_L W_{L-1} \ldots W_{l+1})^+ \neq W^+_{l+1} \ldots W^+_{L-1} W^+_L$, in general! So, $(J^e_{W_l})^+ \neq B_{W_l}$.

However, the recursion on B preserves Moore-Penrose properties 1,2 and 3 for a full-rank, contracting architecture.

Learning rule:

$$\delta_{W_l} = \left(\prod_{i=l+1}^{L} W_i^+\right) e h_{l-1}{}^T$$

Theorem (Levin and Ben-Israel, 2001)

(Informal) Let $f \rightarrow \mathbb{R}^n \rightarrow R^m$ be a vector function we want to minimize, and let T_x be a $\{2\}$-inverse of it's Jacobian matrix, J^f_x. Then, under certain conditions, the iteration

$$x_{k+1} - x_k = -T_{x_k} f(x_k)$$

converges to a fixed point x^* which satisfies

$$T_{x^*} f(x^*) = 0$$

[5]. If T_x is also a $\{1,2,3\}$ inverse, and is full-rank, this solution is in the same nullspace as that reached by gradient descent on $\frac{1}{2}||f||^2$.