A biologically-plausible learning rule using reciprocal feedback connections

Artificial neural networks are frequently used to model neural systems, and have been shown to recapitulate features of biological networks, including in the hippocampus [1] and visual cortex [2] [3]. However, despite it's success in neural modelling, the predominant learning algorithm, backpropagation, has long been considered biologically implausible [4]. One reason is the weight-transport problem - or the problem of how a global error signal can be accurately transmitted across the network to minimize the error resulting from each layer's parameters. Previous solutions to the weight-transport problem have proposed using a separate, error-feedback network to transport an error signal across layers [5]. However, more recent experimental work indicates that cortical feedback connections are more likely to be involved in reconstructing lower-level activity based on activity from higher layers, rather than being exclusively used to transmit top-level error [6] [7] [8]. Under the predictive coding framework, the bidirectional, cortical processing hierarchy is more appropriately modeled as a multi-layered autoencoder [9]. Here, we attempt to unify these two views by showing how autoencoder-like, inverse feedback connections may be used to minimize top-level error in neural networks. Our proposed mechanism, *Reciprocal Feedback*, consists of two contributions: first we show how a modification of the Recirculation autoencoder algorithm [10] is equivalent to learning the Moore-Penrose pseudoinverse. Then, we will show how, using a Newton-like method [11], locally-learned pseudoinverse feedback connections may be used to facilitate an alternative, more biologically-realistic optimization method to gradient descent, by relying on the reciprocal of the forward network rather than its gradient. Overall, we provide a mathematical framework for understanding how the hierarchical, autoencoder-like feedback connections observed in the layers of the cortex may additionally be used as a mechanism for minimizing a global error signal, using only local activity.

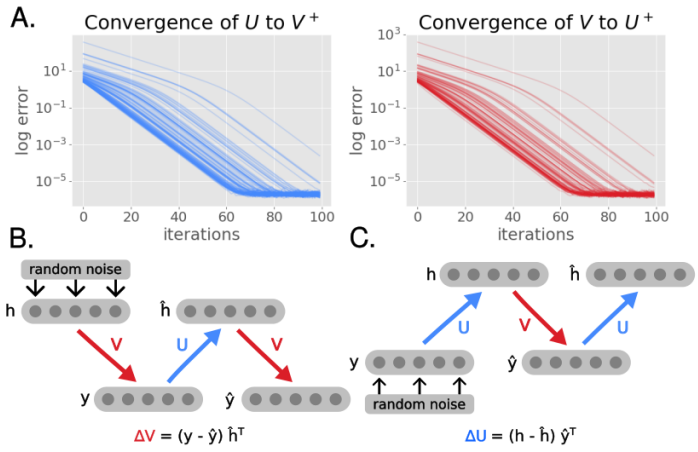

Figure 1: Locally learning the Moore-Penrose pseudoinverse, through a modification of the Recirculation algorithm. **A.** The Frobenius norm between U and V^+ decreases exponentially (and vice versa). In this simulation, the network is linear, and includes decay. U and V were randomly initialized with a low condition number, and trained concurrently. **B.** Learning dynamics of U and V within a single layer, expanded horizontally across time to show recurrence. Referenced from [10].

Local learning of each layer pseudoinverse. First, we show that a modified version of the Recirculation algorithm is capable of learning a pair of pseudoinverse forward (U) and feedback (V) connections, when trained on random, mean-zero noise inputs (y). A theorem described in [12] states that for a low-condition number matrix A, starting from a matrix X_0 satisfying $X_0A = (X_0A)^T$, the sequence generated by:

$$X_{t+1} - X_t = X_t - X_tAX_t$$

converges to A^+ (where A^+ is the pseudoinverse of A). When we remove the nonlinearity from the dynamics of the Recirculation algorithm and average over mean-zero inputs, we find that the learning rules for U and V are nearly identical to the iterative computation of the pseudoinverse (proof omitted). Intuitively, the pseudoinverse can be thought of as the linear, analytical equivalent of an autoencoder feedback connection . The activations of each layer may be approximately reconstructed using the activations of the layer above it - similarly to the dynamics observed experimentally in the cortex. Physically, this learning procedure may be implemented during a "sleep" phase, in which random noise inputs are projected to each layer, so that the forward and feedback connections may align. A similar, biologically-plausible two-phase learning procedure was previously used in the wake-sleep algorithm [13].

Training the whole network. Using each locally-learned, layer-wise pseudoinverse, we now can minimize a global error signal across the whole, multi-layered network. To train the whole network we will be considering a standard, fully-connected, feedforward architecture, with the final-layer error

vector defined as: $e = h_L - h_L{}^*$. We denote each layer's pre-activation vector as $a_l = W_l h_{l-1}$, and its activation vector as $h_l = \sigma(a_l)$. The Jacobian of the error with respect to the layer parameter W_l can be derived using the recursive expression $J^E_{W_l} = (a_{l-1} \otimes J^E_{h_l})$, and the Jacobian with respect to the activation vector h_l as $J^E_{h_l} = J^E_{h_{l+1}} \mathcal{D}_\sigma W_{l+1}$. Informally, the Newton-like method we use [11] states that under certain conditions: if $F : (X_0, B_r(y_0)) \to \mathbb{R}^m$ is a vector function with Jacobian A, with left reciprocal T, then there exists a solution y^* (such that $TF(x, y^*) = 0$) which can be obtained using the iteration: $y_{t+1} = y_t - TF(x, y_t)$. By using the pseudoinverse at each layer, we get recursive expressions for the activation reciprocal: $B_l = W^+_{l+1} \mathcal{D}^+_\sigma B_{l+1}$, and the parameter reciprocal: $\mathcal{B}_l = \left(a^+_{l-1} \otimes B_l\right)$. Using these left reciprocals, we can minimize error by shifting each weight matrix parameter in the direction defined by $\delta^t_{W_l} \propto B_l e(x, W_l) a^T_{l-1}$, resulting in the update rule: $W^{t+1}_l = W^t_l - \lambda \delta^t_{W_l}$ (where λ is a scalar learning rate). Overall, we show how locally-learned, autoencoder-like, pseudoinverse connections can be used to minimize a global error signal, using a Newton-like optimization method - suggesting a biologically-plausible alternative to backpropagation that is more aligned with the structure expected under the predictive coding framework. Computational simulations show a similar asymptotic performance to backpropagation, in fewer iterations than comparable biologically-plausible learning rules, such as Random Feedback Alignment [14].

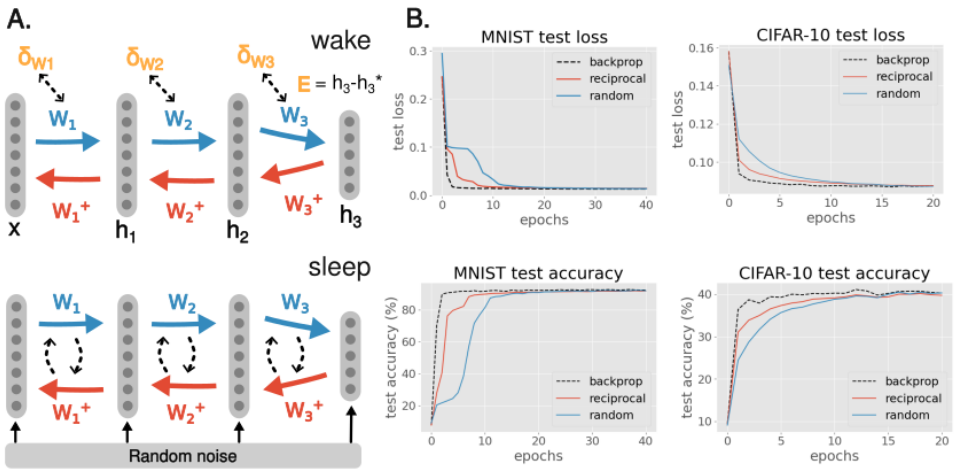

Figure 2: A. Training is split into two phases: a wake phase, where inputs are propagated through the whole network and the global, top-layer error is minimized; and a sleep phase, where forward and feedback weights are aligned to the pseudoinverse of each other. **B.** In-silico implementation on feedforward, fully-connected networks trained on classification tasks. All three learning rules were trained with the same hyperparameters, and reach a similar asymptotic error. Reciprocal feedback converges after fewer iterations than the Random feedback algorithm [14] in both MNIST and CIFAR-10.

Related work. Our work is related to the Target Propagation algorithm [15], in the sense that each layer can be thought of as an autoencoder. However, instead of propagating auxillary "targets" from the top layer, we propagate the error backwards directly. Another biologically-plausible algorithm which utilizes a "wake" and "sleep" phase is the weight-mirror algorithm [16], in which the "sleep" phase learns the weight transpose at each layer - resulting in a closer approximation to backpropagation, but less aligned with the autoencoder-like structure expected. Given the use of pseudoinverses in the Newton-like method, it may be supposed that it is related to Gauss-Newton optimization. However, while Gauss-Newton optimization uses the exact pseudoinverse of the whole network, we use a composition of each layer's pseudoinverse to form a non-unique, left-reciprocal. However, this method is still understudied in the context of neural networks, and may follow a different learning trajectory to gradient descent.

1. Chen, Y. et al. *Neuron* **112,** 2645–2658.e4 (2024). 2. Yamins, D. L. K. et al. *PNAS* **111,** 8619–8624 (May 2014). 3. Khaligh-Razavi, S.-M. et al. *PLOS Comp. Bio.* **10** (ed Diedrichsen, J.) e1003915 (Nov. 2014). 4. Crick, F. *Nature* **337,** 129–132 (Jan. 1989). 5. Lillicrap, T. P. et al. *Nat. Rev. Neuros.* **21,** 335–346 (Apr. 2020). 6. Mumford, D. *Biol. Cybern.* **66,** 241–251 (Jan. 1992). 7. Favila, S. E. et al. *Nat. Comm.* **13** (Oct. 2022). 8. Linde-Domingo, J. et al. *Nat. Comm.* **10** (Jan. 2019). 9. Marino, J. *CoRR* **abs/2011.07464** (2020). 10. Hinton E., G. et al. *AIP* (1988). 11. Hildebrandt, T. H. et al. *Transactions of the AMS* **29,** 127–153 (1927). 12. Ben-Israel, A. et al. *SINUM* **3,** 410–419 (1966). 13. Hinton, G. E. et al. *Science* **268,** 1158–1161 (May 1995). 14. Lillicrap, T. P. et al. *Nat. Comm.* **7** (Nov. 2016). 15. Bengio, Y. *How Auto-Encoders Could Provide Credit Assignment in Deep Networks via Target Propagation* 2014. 16. Akrout, M. et al. *CoRR* **abs/1904.05391** (2019).