

Biologically Motivated Algorithms for Propagating Local Target Representations

Alexander G. Ororbia*

Rochester Institute of Technology
102 Lomb Memorial Drive, Rochester NY, USA 14623

Ankur Mali*

Penn State University
Old Main, State College, PA 16801

Abstract

Finding biologically plausible alternatives to back-propagation of errors is a fundamentally important challenge in artificial neural network research. In this paper, we propose a simple learning algorithm called error-driven Local Representation Alignment (LRA-E), which has strong connections to predictive coding, a theory that offers a mechanistic way of describing neurocomputational machinery. In addition, we propose an improved variant of Difference Target Propagation, another procedure that comes from the same family of algorithms as Local Representation Alignment. We compare our learning procedures to several other biologically-motivated algorithms, including two feedback alignment algorithms and Equilibrium Propagation. In two benchmark datasets, we find that both of our proposed learning algorithms yield stable performance and strong generalization abilities in comparison to other competing back-propagation alternatives when training deeper, highly nonlinear networks, with LRA-E performing the best overall.

Behind the modern achievements in artificial neural network research is back-propagation of errors (Rumelhart, Hinton, and Williams 1988) (or “backprop”), the key training algorithm used in computing updates for the parameters that define the computational architectures applied to problems ranging from computer vision to speech recognition. However, though neural architectures are inspired by our current neuro-scientific understanding of the human brain, the connections to the actual mechanisms that compose systems of natural neurons are often very loose, at best. More importantly, back-propagation of errors faces some of the strongest neuro-biological criticisms, argued to be a highly implausible way in which learning occurs in the human brain.

Among the many problems with back-propagation of errors, some of the most prominent include: 1) the “weight transport problem”, where the feedback weights that carry back error signals must be the transposes of the feedforward weights, 2) forward propagation and backward propagation utilize different computations, and 3) the error gradients are stored separate from the activations. These problems, as originally argued in (Ororbia II et al. 2017; Ororbia et al. 2018), largely center around the one critical component of backprop—the global feedback pathway

needed for transporting error derivatives across the system. This pathway is necessary given the design of modern supervised learning systems—a loss function measures the error between an artificial neural system’s output units and some target (such as a class label) and the global pathway relates how the internal processing elements affect this error. When considering modern theories of the brain (Grossberg 1982; Rao and Ballard 1999; Huang and Rao 2011; Clark 2013), which posit that local computations occur at multiple levels of the somewhat hierarchical structure of natural neural systems, this global pathway should not be necessary to learn effectively. Furthermore, this pathway is the source behind many practical problems that make training more complex, very deep networks difficult—as a result of the many multiplications that underly traversing along this global feedback pathway, error gradients will either explode or vanish (Pascanu, Mikolov, and Bengio 2013). In trying to fix this particular issue, gradients can be kept within reasonable magnitudes by requiring layers to behave sufficiently linearly (which prevents saturation of the post-activation function used, which yield zero gradient). However, this remedy creates other highly undesirable side-effects, such as the well-known problem of adversarial samples (Szegedy et al. 2013; Ororbia II, Kifer, and Giles 2017) and prevents the usage of neuro-biological mechanisms such as lateral competition and discrete-valued/stochastic activation functions (since this pathway requires precise knowledge of the activation function derivatives (Bengio et al. 2015)).

If we remove this global feedback pathway, we create a new problem—what are the learning signals for the hidden processing elements? This problem is one of the main concerns of the recently introduced *Discrepancy Reduction* family of learning algorithms (Ororbia II et al. 2017). In this paper, we will develop two learning algorithms within this family—error-driven Local Representation Alignment and adaptive noise Difference Target Propagation. In experiments on two classification benchmarks, we will show that these two algorithms generalize better than a variety of other biologically motivated learning approaches, all without employing the global feedback pathway required by back-propagation.

Coordinated Local Learning Algorithms

Algorithms within the Discrepancy Reduction (Ororbia II et al. 2017) family offer computational mechanisms to complete

* The first two authors contributed equally.
Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

two key steps when learning from a pattern or mini-batch of patterns. These steps include:

1. Search for latent representations that better explain the input/output, also known as target representations. This creates the need for local (higher-level) objectives that will help guide the current latent representations towards better ones.
2. Reduce, as much as possible, the mismatch between a model’s currently “guessed” representations and target representations. The sum of the internal, local losses is also defined as the total discrepancy in a system, and can also be thought of a sort of pseudo-energy function.

This general process forms the basis of what we call *coordinated local learning rules*. Computing targets with these kinds of rules should not require an actual pathway, as in back-propagation, and instead make use of top-down and bottom-up signals to generate targets. This idea is particularly motivated by the theory of predictive coding (Panichello, Cheung, and Bar 2013) (which started to impact modern machine learning applications (Li and Liu 2018)), which claims that the brain is in a continuous process of creating and updating hypotheses (using error information) to predict the sensory input. This paper will explore two ways in which this hypothesis updating (in the form of local target creation) might happen: 1) through error-correction in Local Representation Alignment (LRA), and 2) through repeated encoding and decoding as in Difference Target Propagation (DTP).

The idea of learning locally, in general, is slowly becoming prominent in the training of artificial neural networks, with recent proposals including decoupled neural interfaces (Jaderberg et al. 2016) and kickback (Balduzzi, Vanchinathan, and Buhmann 2015) (which was derived specifically for regression problems). Furthermore, (Whittington and Bogacz 2017) demonstrated that neural models using simple local Hebbian updates, developed within a predictive coding framework, could efficiently conduct supervised learning. Far earlier approaches that employed local learning included the layer-wise training procedures that were once used to build models for unsupervised learning (Bengio et al. 2007), supervised learning (Lee et al. 2014), and semi-supervised learning (Ororbia II et al. 2015; Ororbia II, Giles, and Reitter 2015). The key problem with these older algorithms is that they were greedy—a model was built from the bottom-up, freezing lower-level parameters as higher-level feature detectors were learnt.

Another important idea underlying algorithms such as LRA and DTP is that learning is possible with asymmetry—which directly resolves the weight-transport problem (Grossberg 1987; Liao, Leibo, and Poggio 2016), another strong neuro-biological criticism of backprop. This is even possible, surprisingly, if the feedback weights are random and fixed, which is at the core of two algorithms we will also compare to—Random Feedback Alignment (RFA) (Lillicrap et al. 2016) and Direct Feedback Alignment (DFA) (Nøkland 2016). RFA replaces the transpose of the feedforward weights in backprop with a similarly-shaped random matrix while DFA directly connects the output layer’s pre-activation derivative to each layer’s post-activation. It was shown in (Ororbia II et al. 2017;

Ororbia et al. 2018) that these feedback loops would be better suited in generating target representations.

Local Representation Alignment

To concretely describe how LRA is practically implemented, we will specify how LRA is applied to a 3-layer feedforward network, or multilayer perceptron (MLP). Note that LRA generalizes to models with more layers ($L \geq 3$).

The pre-activities of the MLP at layer ℓ are denoted as \mathbf{h}^ℓ while the post-activities, or the values output by the non-linearity $\phi_\ell(\cdot)$, are denoted as \mathbf{z}^ℓ . The target variable used to correct the output units (\mathbf{z}^L) is denoted as \mathbf{y}_z^L ($\mathbf{y}_z^L = \mathbf{y}$, or $\mathbf{y}_z^L = \mathbf{x}$ if we are learning an auto-associative function). Connecting one layer of neurons $\mathbf{z}^{\ell-1}$, with pre-activities $\mathbf{h}^{\ell-1}$, to another layer \mathbf{z}^ℓ , with pre-activities \mathbf{h}^ℓ , is a set of synaptic weights W_ℓ . The forward propagation equations for computing pre-activation and post-activation values for a layer ℓ would then simply be:

$$\mathbf{h}^\ell = W_\ell \mathbf{z}^{\ell-1}, \quad \mathbf{z}^\ell = \phi_\ell(\mathbf{h}^\ell) \quad (1)$$

Before computing targets or updates, we first must define the set of local losses, one per layer of neurons except for the input neurons, that constitute the measure of total discrepancy inside the MLP, $\{\mathcal{L}_1(\mathbf{y}_z^1, \mathbf{z}^1), \mathcal{L}_2(\mathbf{y}_z^2, \mathbf{z}^2), \mathcal{L}_3(\mathbf{y}_z^3, \mathbf{z}^3)\}$. With losses defined, we can then explicitly formulate the error units for each layer as well, since any given layer’s error units correspond to the first derivative of that layer’s loss with respect to that layer’s post-activation values. For the MLP’s output layer, we could assume a categorical distribution, which is appropriate for 1-of- k classification tasks, and use the following negative log likelihood loss:

$$\mathcal{L}_\ell(\mathbf{y}_z^\ell, \mathbf{z}^\ell) = -\frac{1}{2} \sum_{i=1}^{|\mathbf{z}|} \mathbf{y}_z^\ell[i] \log \mathbf{z}^\ell[i],$$

$$\mathbf{e}_\ell = \mathbf{e}_\ell(\mathbf{y}_z^\ell, \mathbf{z}^\ell) = \frac{-\mathbf{y}_z^\ell}{\mathbf{z}^\ell}, \quad (2)$$

where the loss is computed over all dimensions $|\mathbf{z}|$ of the vector \mathbf{z} (where a dimension is indexed/accessed by integer i). Note that for this loss function, we assume that \mathbf{z} is a vector of probabilities computed by using the softmax function as the output nonlinearity, $\mathbf{z}^3 = \frac{\exp(\mathbf{h}^3)}{\sum_i \exp(\mathbf{h}_i^3)}$. For the hidden layers, we can choose between a wider variety of loss functions, and in this paper, we experimented with assuming either a Gaussian or Cauchy distribution over the hidden units. For the Gaussian distribution (or L2 norm), we have the following loss and error unit pair:

$$\mathcal{L}_\ell(\mathbf{z}, \mathbf{y}) = \frac{1}{(2\sigma^2)} \sum_{i=1}^{|\mathbf{z}|} (\mathbf{y}_i - \mathbf{z}_i)^2$$

$$\mathbf{e}_\ell = \mathbf{e}_\ell(\mathbf{y}_z^\ell, \mathbf{z}^\ell) = -(2\sigma^2)(\mathbf{y}_z^\ell - \mathbf{z}^\ell) \quad (3)$$

where σ^2 represents fixed scalar variance (we set $\sigma^2 = 1$).

For the Cauchy distribution (or log-penalty), we obtain:

$$\mathcal{L}_\ell(\mathbf{y}, \mathbf{z}) = \sum_{i=1}^{|\mathbf{z}|} \log(1 + (\mathbf{y}_i - \mathbf{z}_i)^2)$$

$$\mathbf{e}_\ell = \mathbf{e}_\ell(\mathbf{y}_z^\ell, \mathbf{z}^\ell) = \frac{-2(\mathbf{y}_z^\ell - \mathbf{z}^\ell)}{(1 + (\mathbf{y}_z^\ell - \mathbf{z}^\ell)^2)}. \quad (4)$$

For the activation function used in calculating the hidden post-activities, we use the hyperbolic tangent, or $\phi_\ell(v) = \frac{\exp(2v)-1}{\exp(2v)+1}$. Using the Cauchy distribution proved particularly useful in our experiments, most likely because it encourages sparse representations, which aligns nicely with the biological considerations of sparse coding (Olshausen and Field 1997) and predictive sparse decomposition (Kavukcuoglu, Ranzato, and LeCun 2010) as well as the lateral competition (Rao and Ballard 1999) that naturally occurs in groups of neural processing elements. These are relatively simple local losses that one can use to measure the agreement between the representation and target and future work should entail developing even better metrics.

Algorithm 1 LRA-E: Target computation.

Input: sample (\mathbf{y}, \mathbf{x}) and $\Theta = \{W_1, W_2, W_3, E_2, E_3\}$
 // Procedure for computing error units & targets
function COMPUTETARGETS $((\mathbf{y}, \mathbf{x}), \Theta)$
 // Run feedforward weights to get activities
 $\mathbf{h}^1 = W_1 \mathbf{z}^0, \mathbf{z}^1 = \phi_1(\mathbf{h}^1)$
 $\mathbf{h}^2 = W_2 \mathbf{z}^1, \mathbf{z}^2 = \phi_2(\mathbf{h}^2)$
 $\mathbf{h}^3 = W_3 \mathbf{z}^2, \mathbf{z}^3 = \phi_3(\mathbf{h}^3)$
 $\mathbf{y}_z^3 \leftarrow \mathbf{y}$
 $\mathbf{e}_3 = \frac{-\mathbf{y}_z^3}{\mathbf{z}^3}, \mathbf{y}_z^2 \leftarrow \phi_2(\mathbf{h}^2 - \beta(E_3 \mathbf{e}_3))$
 $\mathbf{e}_2 = -2(\mathbf{y}_z^2 - \mathbf{z}^2)$
 $\mathbf{y}_z^1 \leftarrow \phi_1(\mathbf{h}^1 - \beta(E_2 \mathbf{e}_2))$
 $\mathbf{e}_1 = -2(\mathbf{y}_z^1 - \mathbf{z}^1)$
 $\Lambda = (\mathbf{z}^3, \mathbf{z}^2, \mathbf{z}^1, \mathbf{h}^3, \mathbf{h}^2, \mathbf{h}^1, \mathbf{e}^3, \mathbf{e}^2, \mathbf{e}^1)$
Return Λ

Algorithm 2 LRA-E: Update computation.

Input: sample (\mathbf{y}, \mathbf{x}) and calculations Λ
 // Procedure for computing weight updates
function CALCUPDATES $((\mathbf{y}, \mathbf{x}), \Theta, \Lambda)$
 $\Delta W_3 = (\mathbf{e}_3 \otimes \phi'_3(\mathbf{h}_3))(\mathbf{z}_2)^T$
 $\Delta W_2 = (\mathbf{e}_2 \otimes \phi'_2(\mathbf{h}_2))(\mathbf{z}_1)^T$
 $\Delta W_1 = (\mathbf{e}_1 \otimes \phi'_1(\mathbf{h}_1))(\mathbf{x})^T$
 $\Delta E_3 = -\gamma(\Delta W_3)^T$
 $\Delta E_2 = -\gamma(\Delta W_2)^T$
Return $(\Delta W_3, \Delta W_2, \Delta W_1, \Delta E_3, \Delta E_2)$
function CALCUPDATES $((\mathbf{y}, \mathbf{x}), \Theta, \Lambda)$
 $\Delta W_3 = \mathbf{e}_3(\mathbf{z}_2)^T$
 $\Delta W_2 = \mathbf{e}_2(\mathbf{z}_1)^T$
 $\Delta W_1 = \mathbf{e}_1(\mathbf{x})^T$
 $\Delta E_3 = -\gamma(\Delta W_3)^T$
 $\Delta E_2 = -\gamma(\Delta W_2)^T$
Return $(\Delta W_3, \Delta W_2, \Delta W_1, \Delta E_3, \Delta E_2)$

With local losses specified and error units implemented, all that remains is to define how targets are computed and what the parameter updates will be. At any given layer \mathbf{z}^ℓ , starting at the output units (in our example, \mathbf{z}^3), we calculate the target for the layer below $\mathbf{z}^{\ell-1}$ by multiplying the error unit values at ℓ by a set of synaptic error weights E_ℓ . This projected displacement, weighted by the modulation factor β ,¹ is then subtracted from the initially found pre-activation of the layer below $\mathbf{h}^{\ell-1}$. This updated pre-activity is then run through the appropriate nonlinearity to calculate the final target $\mathbf{y}_z^{\ell-1}$. This computation amounts to:

$$\mathbf{e}_\ell = -2(\mathbf{y}_z^\ell - \mathbf{z}^\ell), \quad \Delta \mathbf{h}^{\ell-1} = E_\ell \mathbf{e}_\ell \quad (5)$$

$$\mathbf{y}_z^{\ell-1} \leftarrow \phi_{\ell-1}(\mathbf{h}^{\ell-1} - \beta(\Delta \mathbf{h}^{\ell-1})). \quad (6)$$

Once the targets for each layer have been found, we can then use the local loss $\mathcal{L}^\ell(\mathbf{y}_z^\ell, \mathbf{z}^\ell)$ to compute updates to the weights W_ℓ and its corresponding error weights E_ℓ .² The update calculation for parameters at layer ℓ would be:

$$\Delta W_\ell = (\mathbf{e}_\ell \otimes \phi'_\ell(\mathbf{h}_\ell))(\mathbf{z}_{\ell-1})^T \ \& \ \Delta E_\ell = -\gamma(\Delta W_\ell)^T, \quad (7)$$

or,

$$\Delta W_\ell = \mathbf{e}_\ell(\mathbf{z}_{\ell-1})^T \ \& \ \Delta E_\ell = \gamma(\Delta W_\ell)^T \quad (8)$$

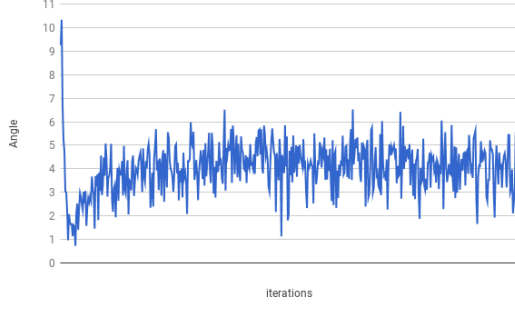
where \otimes indicates the Hadamard product and γ is a decay factor (a value that we found should be set to less than 1.0) meant to ensure that the error weights change more slowly than the forward weights. Note that the second variation of the update rule does not require $\phi'(\cdot)$, which makes it particularly attractive in that it does not require the first derivative of the activation function thus permitting the use of discrete and stochastic operations. The update for each set of error weights is simply proportional to the negative transpose of the update computed for its matching forward weights, which is a computationally fast and cheap rule we propose inspired by (Rao and Ballard 1997).

In Algorithms 1 and 2, we show how the equations in this section combine together to create the full procedure for training a 3-layer MLP, assuming Gaussian local loss functions and their respective error units. The model is defined by parameters $\Theta = \{W^1, W^2, W^3, E^2, E^3\}$ (biases \mathbf{c}^ℓ are omitted for clarity). We will refer to this algorithm as *LRA-E*.

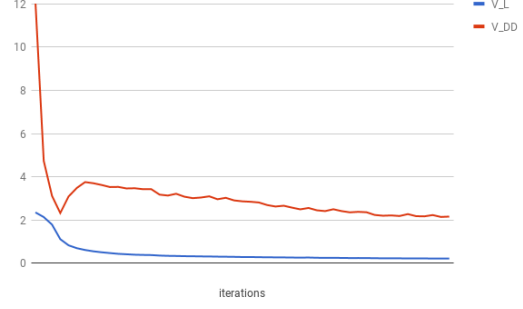
With a local loss assigned to each hidden layer, we can measure our neural model's total internal discrepancy, $D(\mathbf{y}, \mathbf{x})$, for a given data point as a simple linear combination of all of the internal local losses. Figure 1(b) shows the 3-layer MLP example developed in this section (256 units each), trained by stochastic gradient descent (SGD) and mini-batches of 50 image samples, over the first 20 epochs of learning using a categorical output loss and two Gaussian local losses. While the output loss (V_L) continues to decrease, the total discrepancy (V_{DD}) does not always appear to do so, especially in

¹In the experiments of this paper, a value of $\beta = 0.1$, found with only minor tuning in preliminary experiments on a subset of training and validation data proved to be effective in general.

²Except for very bottom set of forward weights, W_1 , of which there are no error corresponding error weights.



(a) Gradient angle between LRA-E and backprop.



(b) Validation NLL and total discrepancy .

Figure 1: In Figure 1(a), we compare the updates calculated by LRA-E and backprop. In Figure 1(b), we show how the total discrepancy, as measured in an LRA-trained MLP, evolves during training, alongside the output loss.

the earlier part of learning. However, since each layer will try to minimize the mismatch between itself and a target value, any fluxes, or local loss values that actually increases instead of decreases which might raise the total discrepancy, will be taken care of later as the model starts generating better targets. So long as the angle of the updates computed from LRA are within 90 degrees of the updates obtained by back-propagation of errors, LRA will move parameters towards the same general direction as back-propagation, which greedily points to the direction of steepest descent, and still find good local optima. In Figure 1(a), this does indeed appear to be the case—for the same 3-layer MLP trained for Figure 1(b) in this section, we compare the updates calculated by *LRA-E* with those given by back-propagation after each mini-batch. The angle, fortunately, while certainly non-zero, never deviates too far from the direction pointed by back-propagation (at most 11 degrees) and remains relatively stable throughout the learning process.

Improving Difference Target Propagation

As mentioned earlier, Difference Target Propagation (DTP) (and also, less directly, recirculation (Hinton and McClelland 1988; O’Reilly 1996)), like LRA-E, also falls under the same family of algorithms concerned with minimizing internal discrepancy, as shown in (Ororbias II et al. 2017; Ororbias et al. 2018). However, DTP takes a very different approach to computing alignment targets—instead of transmitting messages through error units and error feedback weights as in LRA-E, DTP employs feedback weights to learn the inverse of the mapping created by the feedforward weights. However, (Ororbias et al. 2018) showed that DTP struggles to assign good local targets as the network becomes deeper, i.e., more highly nonlinear, facing an initially promising albeit brief phase in learning where generalization error decreases (within the first few epochs) before ultimately collapsing (unless very specific initializations are used). One potential cause of this failure could be the lack of a strong enough mechanism to globally coordinate the local learning problems created by the encoder-decoder pairs that underlie the system. In particular, we hypothesize that this problem might be coming

from the noise injection scheme, which is local and fixed, offering no adaptation to each specific layer and making some of the layerwise optimization problems more difficult than necessary. Here, we will aim to remove this potential cause through an adaptive layerwise corruption scheme.

Assuming we have a target calculated from above \mathbf{y}_z^ℓ , we consider the forward weights W_ℓ connecting the layer $\mathbf{z}^{\ell-1}$ to layer \mathbf{z}^ℓ and the decoding weights E_ℓ that define the inverse mapping between the two. The first forward propagation step is the same as in Equation 1. In contrast to LRA-E’s error-driven way of computing targets, we consider each pair of neuronal layers, $(\mathbf{z}^\ell, \mathbf{z}^{\ell-1})$, as forming a particular type of encoding/decoding cycle that will be used in computing layerwise targets. To calculate the target $\mathbf{y}_z^{\ell-1}$, we update the original post-activation $\mathbf{z}^{\ell-1}$ using the linear combination of two applications of the decoding weights as follows:

$$\mathbf{y}_z^{\ell-1} = \mathbf{z}^{\ell-1} - (\phi_{\ell-1}(V_\ell \mathbf{z}^\ell) + \phi_{\ell-1}(V_\ell \mathbf{y}_z^\ell)) \quad (9)$$

where we see that we decode two times, one from the original post-activation calculated from the feedforward pass of the MLP and another from the target value generated from the encoding/decoding process from the layer pair above, e.g. $(\mathbf{z}^{\ell+1}, \mathbf{z}^\ell)$. This will serve as the target when training the forward weights for the layer below $W_{\ell-1}$. To train the inverse-mapping weights V_ℓ , as required by the original proposed version of DTP, zero-mean Gaussian noise, $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with fixed standard deviation σ , is injected into $\mathbf{z}^{\ell-1}$ followed by re-running the encoder and the decoder on this newly corrupted activation vector. Formally, this is defined as:

$$\hat{\mathbf{y}}_z^{\ell-1} = \mathbf{z}^{\ell-1} + \epsilon, \quad \hat{\mathbf{z}}^{\ell-1} = \phi_{\ell-1}(V_\ell \phi_\ell(W_\ell \hat{\mathbf{y}}_z^{\ell-1})) \quad (10)$$

This process we will refer to as *DTP*. In our proposed, improved variation of DTP, or *DTP- σ* , we will take an “adaptive” approach to the noise injection process ϵ . To develop our adaptive noise scheme, we have taken some insights from studies of biological neuron systems, which show there are varying levels of signal corruption in different neuronal layers/groups (D. and Yngve ; Tomko and Crapper 1974; Tolhurst, Movshon, and Dean 1983; Shadlen and Newsome 1998). It has been argued that this noise variability enhances

neurons’ overall ability to detect and transmit signals across a system (Shu et al. 2003; Kruglikov and Dertinger 1994; Shadlen and Newsome 1998) and, furthermore, that the presence of this noise yields more robust representations (Cordo et al. 1996; Shadlen and Newsome 1998; Faisal, Selen, and Wolpert 2008). There also is biological evidence demonstrating that an increase in the noise level across successive groups of neurons is thought to help in local neural computation (Shadlen and Newsome 1998; Sarpeshkar 1998; Laughlin, de Ruyter van Steveninck, and Anderson 1998).

In light of this, the standard deviation σ of the noise process should be a function of the noise across layers, and an interesting way in which we implemented this was to make σ_ℓ (the standard deviation of the noise injection at layer ℓ) a function of the local loss measurements. At the top layer, we can set the standard deviation to small, fixed value $\sigma_L = \alpha$ ($\alpha = 0.01$ worked well in our experiments). The standard deviation for the layers below would be a function of where the noise process is within the network, indexed by ℓ . This means that:

$$\sigma_\ell = \sigma_{\ell+1} - \mathcal{L}_\ell(\mathbf{y}_z^{\ell-1}, \mathbf{z}^{\ell-1}) \quad (11)$$

noting that the local loss chosen for DTP is a Gaussian loss (but with the input arguments flipped—the target value is now the corrupted initial encoding and the prediction is the clean, original encoding, or $\mathcal{L}_\ell(\mathbf{z} = \mathbf{y}_z^{\ell-1}, \mathbf{y} = \mathbf{z}^{\ell-1})$).

The updates to the weights are calculated by differentiating each local loss with respect to the appropriate encoder weights, or $\Delta W_{\ell-1} = \frac{\partial \mathcal{L}(\mathbf{z}^{\ell-1}, \mathbf{y}_z^{\ell-1})}{\partial W_{\ell-1}}$, or with respect to the decoder synaptic weights $\Delta V_\ell = \frac{\partial \mathcal{L}(\mathbf{z}^\ell, \hat{\mathbf{y}}_z^\ell)}{\partial V_\ell}$. Note that the order of the input arguments to each loss function for these two partial derivatives is important for obtaining the correct sign to multiply the gradients by, and furthermore staying aligned with the original formulation of DTP (Lee et al. 2015a), .

As we will see in our experimental results, *DTP- σ* is a much more stable learning algorithm, especially when training deeper/wider networks. *DTP- σ* benefits from a stronger form of overall coordination among its internal encoding/decoding sub-problems through the pair-wise comparison of local loss values that drive the hidden layer corruption.

A Comment on the Efficiency of LRA-E and DTP

It should be noted that *LRA-E* is faster than *DTP* and *DTP- σ* in calculating targets. Specifically, if we focus on matrix multiplications used to find updates, which would take up the bulk of the computation underlying both processes, *LRA-E* only requires $2(L-1)$ matrix multiplications while *DTP* and *DTP- σ* require $4(L-3) + L$ multiplications. In particular, the bulk of DTP’s primary expense comes from its approach to computing the targets for the hidden layers—it requires 2 applications of the encoder parameters (1 of these is from the initial feedforward pass through the network) and 3 applications of the decoder parameters in order to generate targets to train the forward weights and the inverse-mapping weights.

Experimental Results

In this section, we present experimental results of training MLPs using a variety of learning algorithms.

MNIST: This dataset³ contains 28×28 images with gray-scale pixel feature values in the range of $[0, 255]$. The only preprocessing applied to this data is to normalize the pixel values to the range of $[0, 1]$ by dividing them by 255.

Fashion MNIST: This database (Xiao, Rasul, and Vollgraf 2017) contains 28×28 grey-scale images of clothing items, meant to serve as a much more difficult drop-in replacement for MNIST itself. Training contains 60000 samples and testing contains 10000, each image is associated with one of 10 classes. We create a validation set of 2000 samples from the training split. Preprocessing was the same as on MNIST.

For both datasets and all models, over 100 epochs, we calculate updates over mini-batches of 50 samples. Furthermore, we do not regularize parameters any further, such as through drop-out (Srivastava et al. 2014) or penalties placed on the weights. All feedforward architectures for all experiments were of either 3, 5, or 8 hidden layers of 256 processing elements. The post-activation function used was the hyperbolic tangent and the top layer was chosen to be a maximum-entropy classifier (employing the softmax function). The output layer objective for all algorithms was to minimize the categorical negative log likelihood.

Parameters were initialized using a scheme that gave best performance on the validation split of each dataset on a per-algorithm basis. Though we wanted to use very simple initialization schemes for all algorithms, in preliminary experiments, we found that the feedback alignment algorithms as well as *DTP* (and *DTP- σ*) worked best when using a uniform fan-in-fan-out scheme (Glorot and Bengio 2010). (Ororbia et al. 2018) confirms this result, originally showing how these algorithms often are unstable or fail to perform well using an initialization based on the uniform or Gaussian distributions. For LRA-E, however, we initialized the parameters using a zero-mean Gaussian distribution, with variance of 0.05.

The choice of parameter update rule was also somewhat dependent on the learning algorithm employed. Again, as shown in (Ororbia et al. 2018), it is difficult to get good, stable performance from algorithms, such as the original DTP, when using simple SGD. As done in (Lee et al. 2015b), we used the RMSprop (Tieleman and Hinton 2012) adaptive learning rate with a global step size of $\lambda = 0.001$. For Backprop, RFA, DFA, and LRA-E, we were able to use SGD ($\lambda = 0.01$).

Classification Performance

In this experiment, we compare all of the algorithms discussed earlier. These include back-propagation of errors (Backprop), Random Feedback Alignment (RFA) (Lillicrap et al. 2014), Direct Feedback Alignment (DFA) (Nøkland 2016), Equilibrium Propagation (Scellier and Bengio 2017) (Equil-Prop) and the original Difference Target Propagation (Lee et al. 2015a) (DTP). Our algorithms include our proposed, improved version of DTP (*DTP- σ*) and the proposed error-driven Local Representation Alignment (LRA-E).

The results of our experiments are presented in Tables 1 and 2. Test and training scores are reported for the set of model parameters that had lowest validation error. Observe that LRA-E is the most stable and consistently well-

³Available at the URL: <http://yann.lecun.com/exdb/mnist/>.

Table 1: MNIST supervised classification results.

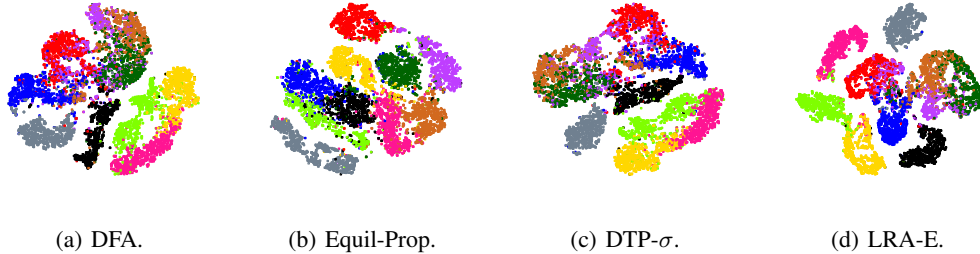
Model	3 Layers		5 Layers		8 Layers	
	Train	Test	Train	Test	Train	Test
<i>Backprop</i>	1.78	3.02	2.4	2.98	2.91	3.02
<i>Equil-Prop</i>	3.82	4.99	7.59	9.21	89.96	90.91
<i>RFA</i>	3.01	3.13	2.99	3.4	3.59	3.76
<i>DFA</i>	4.07	4.17	3.71	3.88	3.81	3.85
<i>DTP</i>	0.74	2.8	4.408	4.94	10.89	10.1
<i>DTP-σ</i> (ours)	0.00	2.38	0.00	2.57	0.00	2.56
<i>LRA-E</i> (ours)	0.86	2.20	0.16	1.97	0.08	2.55

Table 2: Fashion MNIST supervised classification results.

Model	3 Layers		5 Layers		8 Layers	
	Train	Test	Train	Test	Train	Test
<i>Backprop</i>	12.08	14.89	12.1	12.98	11.55	13.21
<i>Equil-Prop</i>	14.72	14.01	16.56	20.97	90.12	89.78
<i>RFA</i>	11.99	12.74	12.09	12.89	12.03	12.71
<i>DFA</i>	13.04	13.41	12.58	13.09	11.59	13.01
<i>DTP</i>	13.6	15.03	21.078	19.66	21.838	17.58
<i>DTP-σ</i> (ours)	7.5	13.95	6.34	12.99	6.5	13.01
<i>LRA-E</i> (ours)	11.25	13.51	9.84	12.31	9.74	12.69

Table 3: Effect of update rule on LRA when training a 3-layer MLP on MNIST.

Model	SGD		Adam		RMSprop	
	Train	Test	Train	Test	Train	Test
<i>LRA, MNIST</i>	0.86	2.20	0.00	1.75	0.69	2.02
<i>LRA, Fashion MNIST</i>	11.25	13.51	5.38	12.42	12.67	14.90

Figure 2: Visualization of the topmost hidden layer extracted from a 5-layer MLP trained by Direct Feedback Alignment (DFA), Equilibrium Propagation (Equil-Prop), adaptive noise Difference Target Propagation (DTP- σ), and error-driven Local Representation Alignment (LRA-E).

performing algorithm compared to the other biologically-motivated backprop alternatives, closely followed by *DTP- σ* . More importantly note algorithms like Equil-Prop and DTP appear to break down when training deeper networks, such as the 8-layer MLP. Note that while DTP was used to successfully train a 7-layer network of 240 units (using RMSprop), we followed the same settings reported for deeper 7 layers network and in our experiments uncovered that the algorithm begins to struggle as the layers are made wider, starting even as soon at the width of 256 we experimented with in this paper. However, this problem is rectified using our variant of

DTP, leading to much more stable performance and even in cases where the algorithm completely overfits the training set (as in the case of 3 and 5 layers for MNIST). Nonetheless, LRA-E still performs the best with respect to generalization across both datasets, despite using a vastly simpler parameter update rule and a naïve initialization scheme. Table 3 shows the results of using update rules other than SGD for LRA-E, e.g., Adam (Kingma and Ba 2014) or RMSprop (Tieleman and Hinton 2012) for a 3-layer MLP, (global step size 0.001 for both algorithms). We see that LRA-E is not only compatible with other learning rate schemes but reaches better

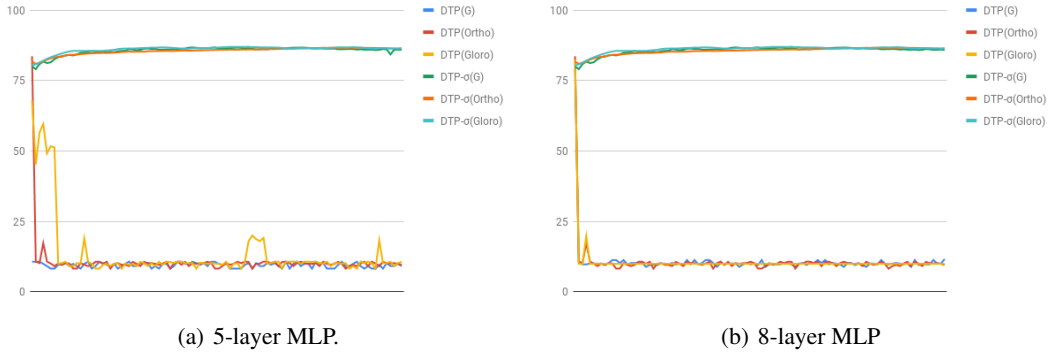


Figure 3: Validation accuracy of DTP vs improved $DTP\text{-}\sigma$ as a function of epoch.

generalization performance when using them, e.g. Adam.

Figure 2 displays a t-SNE (Maaten and Hinton 2008) visualization of the top-most hidden layer of a learned 5-layer MLP using either DFA, Equil-Prop, $DTP\text{-}\sigma$, and LRA-E on Fashion MNIST samples. Qualitatively, we see that all learning algorithms extract representations that separate out the data points reasonably well, at least in the sense that points are clustered based on clothing type. However, it appears that $LRA\text{-}E$ representations yields more strongly separated clusters as evidenced by somewhat wider gaps between them, especially around the pink, blue, and black colored clusters.

Finally, DTP , as also mentioned in (Ororbia et al. 2018), appears to be quite sensitive to its initialization scheme. For both MNIST and Fashion MNIST, we trained DTP and our proposed $DTP\text{-}\sigma$ with three different settings, including random orthogonal, fan-in-fan-out, and simple zero-mean Gaussian initializations. Figure 3 shows the validation accuracy curves of the DTP and $DTP\text{-}\sigma$ as a function of epoch for 5 and 8 layer networks with various weight initializations such as Gaussian (G), Orthogonal (Ortho), and Xavier/Glorot (Gloro). As shown in the figure DTP is highly unstable as the network gets deeper while $DTP\text{-}\sigma$ appears to be less dependent on the weight initialization scheme. Thus, our experiments show some promising evidence of $DTP\text{-}\sigma$ ’s generalization improvement over the original DTP . Moreso, as the test performance indicates in Tables 1 and 2, $DTP\text{-}\sigma$ can, overall, perform nearly as well as $LRA\text{-}E$.

Conclusions

In this paper, we proposed two learning algorithms: error-driven Local Representation Alignment and adaptive noise Difference Target Propagation. On two classification benchmarks, we show strong positive results when training deep multilayer perceptrons. Future work will include investigating how these algorithms fare in the face of much larger-scale tasks and adapting them to problems where labeled data is scarce or the data has a temporal dimension to it.

References

[Balduzzi, Vanchinathan, and Buhmann 2015] Balduzzi, D.; Vanchinathan, H.; and Buhmann, J. M. 2015. Kickback cuts

backprop’s red-tape: Biologically plausible credit assignment in neural networks. In *AAAI*, 485–491.

[Bengio et al. 2007] Bengio, Y.; Lamblin, P.; Popovici, D.; Larochelle, H.; et al. 2007. Greedy layer-wise training of deep networks. *Advances in neural information processing systems* 19:153.

[Bengio et al. 2015] Bengio, Y.; Lee, D.-H.; Bornschein, J.; Mesnard, T.; and Lin, Z. 2015. Towards biologically plausible deep learning. *arXiv preprint arXiv:1502.04156*.

[Clark 2013] Clark, A. 2013. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36(3):181–204.

[Cordo et al. 1996] Cordo, P.; Inglis, J. T.; Verschueren, S.; Collins, J. J.; Merfeld, D. M.; Rosenblum, S.; Buckley, S.; and Moss, F. 1996. Noise in human muscle spindles. *Nature* 383(6603):769–770.

[D. and Yngve] D., A. E., and Yngve, Z. The impulses produced by sensory nerve-endings. *The Journal of Physiology* 61(2):151–171.

[Faisal, Selen, and Wolpert 2008] Faisal, A. A.; Selen, L. P.; and Wolpert, D. M. 2008. Noise in the nervous system. *Nat. Rev. Neurosci.* 9(4):292–303.

[Glorot and Bengio 2010] Glorot, X., and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 249–256.

[Grossberg 1982] Grossberg, S. 1982. How does a brain build a cognitive code? In *Studies of mind and brain*. Springer. 1–52.

[Grossberg 1987] Grossberg, S. 1987. Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science* 11(1):23 – 63.

[Hinton and McClelland 1988] Hinton, G. E., and McClelland, J. L. 1988. Learning representations by recirculation. In *Neural information processing systems*, 358–366.

[Huang and Rao 2011] Huang, Y., and Rao, R. P. 2011. Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science* 2(5):580–593.

- [Jaderberg et al. 2016] Jaderberg, M.; Czarnecki, W. M.; Osindero, S.; Vinyals, O.; Graves, A.; and Kavukcuoglu, K. 2016. Decoupled neural interfaces using synthetic gradients. *arXiv preprint arXiv:1608.05343*.
- [Kavukcuoglu, Ranzato, and LeCun 2010] Kavukcuoglu, K.; Ranzato, M.; and LeCun, Y. 2010. Fast inference in sparse coding algorithms with applications to object recognition. *arXiv preprint arXiv:1010.3467*.
- [Kingma and Ba 2014] Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- [Kruglikov and Dertinger 1994] Kruglikov, I. L., and Dertinger, H. 1994. Stochastic resonance as a possible mechanism of amplification of weak electric signals in living cells. *Bioelectromagnetics* 15(6):539–547.
- [Laughlin, de Ruyter van Steveninck, and Anderson 1998] Laughlin, S. B.; de Ruyter van Steveninck, R. R.; and Anderson, J. C. 1998. The metabolic cost of neural information. *Nat. Neurosci.* 1(1):36–41.
- [Lee et al. 2014] Lee, C.-Y.; Xie, S.; Gallagher, P.; Zhang, Z.; and Tu, Z. 2014. Deeply-Supervised Nets. *arXiv:1409.5185 [cs, stat]*.
- [Lee et al. 2015a] Lee, D.-H.; Zhang, S.; Fischer, A.; and Bengio, Y. 2015a. Difference target propagation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 498–515. Springer.
- [Lee et al. 2015b] Lee, D.-H.; Zhang, S.; Fischer, A.; and Bengio, Y. 2015b. Difference target propagation. In *Proceedings of the 2015th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I, ECMLPKDD'15*, 498–515. Switzerland: Springer.
- [Li and Liu 2018] Li, J., and Liu, H. 2018. Predictive coding machine for compressed sensing and image denoising. In *AAAI*.
- [Liao, Leibo, and Poggio 2016] Liao, Q.; Leibo, J. Z.; and Poggio, T. A. 2016. How important is weight symmetry in backpropagation? In *AAAI*, 1837–1844.
- [Lillicrap et al. 2014] Lillicrap, T. P.; Cownden, D.; Tweed, D. B.; and Akerman, C. J. 2014. Random feedback weights support learning in deep neural networks. *arXiv preprint arXiv:1411.0247*.
- [Lillicrap et al. 2016] Lillicrap, T. P.; Cownden, D.; Tweed, D. B.; and Akerman, C. J. 2016. Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications* 7:13276.
- [Maaten and Hinton 2008] Maaten, L. v. d., and Hinton, G. 2008. Visualizing data using t-sne. *Journal of machine learning research* 9(Nov):2579–2605.
- [Nøkland 2016] Nøkland, A. 2016. Direct feedback alignment provides learning in deep neural networks. In *Advances in Neural Information Processing Systems*, 1037–1045.
- [Olshausen and Field 1997] Olshausen, B. A., and Field, D. J. 1997. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research* 37(23):3311–3325.
- [O'Reilly 1996] O'Reilly, R. C. 1996. Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural computation* 8(5):895–938.
- [Ororbial et al. 2018] Ororbial, A. G.; Mali, A.; Kifer, D.; and Giles, C. L. 2018. Conducting credit assignment by aligning local representations. *arXiv preprint arXiv:1803.01834*.
- [Ororbial II et al. 2015] Ororbial II, A. G.; Reitter, D.; Wu, J.; and Giles, C. L. 2015. Online learning of deep hybrid architectures for semi-supervised categorization. In *Machine Learning and Knowledge Discovery in Databases (Proceedings, ECML PKDD 2015)*, volume 9284 of *Lecture Notes in Computer Science*. Porto, Portugal: Springer. 516–532.
- [Ororbial II et al. 2017] Ororbial II, A. G.; Haffner, P.; Reitter, D.; and Giles, C. L. 2017. Learning to adapt by minimizing discrepancy. *arXiv preprint arXiv:1711.11542*.
- [Ororbial II, Giles, and Reitter 2015] Ororbial II, A. G.; Giles, C. L.; and Reitter, D. 2015. Online semi-supervised learning with deep hybrid boltzmann machines and denoising autoencoders. *arXiv preprint arXiv:1511.06964*.
- [Ororbial II, Kifer, and Giles 2017] Ororbial II, A. G.; Kifer, D.; and Giles, C. L. 2017. Unifying adversarial training algorithms with data gradient regularization. *Neural computation* 29(4):867–887.
- [Panichello, Cheung, and Bar 2013] Panichello, M.; Cheung, O.; and Bar, M. 2013. Predictive feedback and conscious visual experience. *Frontiers in Psychology* 3:620.
- [Pascanu, Mikolov, and Bengio 2013] Pascanu, R.; Mikolov, T.; and Bengio, Y. 2013. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, 1310–1318.
- [Rao and Ballard 1997] Rao, R. P., and Ballard, D. H. 1997. Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural computation* 9(4):721–763.
- [Rao and Ballard 1999] Rao, R. P., and Ballard, D. H. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience* 2(1).
- [Rumelhart, Hinton, and Williams 1988] Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1988. Neurocomputing: Foundations of research. Cambridge, MA, USA: MIT Press. chapter Learning Representations by Back-propagating Errors, 696–699.
- [Sarpeshkar 1998] Sarpeshkar, R. 1998. Analog versus digital: extrapolating from electronics to neurobiology. *Neural Comput* 10(7):1601–1638.
- [Scellier and Bengio 2017] Scellier, B., and Bengio, Y. 2017. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in computational neuroscience* 11:24.
- [Shadlen and Newsome 1998] Shadlen, M. N., and Newsome, W. T. 1998. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *J. Neurosci.* 18(10):3870–3896.
- [Shu et al. 2003] Shu, Y.; Hasenstaub, A.; Badoual, M.; Bal, T.; and McCormick, D. A. 2003. Barrages of synaptic

activity control the gain and sensitivity of cortical neurons. *J. Neurosci.* 23(32):10388–10401.

[Srivastava et al. 2014] Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15(1):1929–1958.

[Szegedy et al. 2013] Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

[Tieleman and Hinton 2012] Tieleman, T., and Hinton, G. 2012. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning.

[Tolhurst, Movshon, and Dean 1983] Tolhurst, D.; Movshon, J.; and Dean, A. 1983. The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Research* 23(8):775 – 785.

[Tomko and Crapper 1974] Tomko, G. J., and Crapper, D. R. 1974. Neuronal variability: non-stationary responses to identical visual stimuli. *Brain Research* 79(3):405 – 418.

[Whittington and Bogacz 2017] Whittington, J. C. R., and Bogacz, R. 2017. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural Computation* 29(5):1229–1262. PMID: 28333583.

[Xiao, Rasul, and Vollgraf 2017] Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.