Learning without feedback: Fixed random learning signals allow for feedforward training of deep neural networks

Charlotte Frenkel*‡, Martin Lefebvre‡ and David Bol

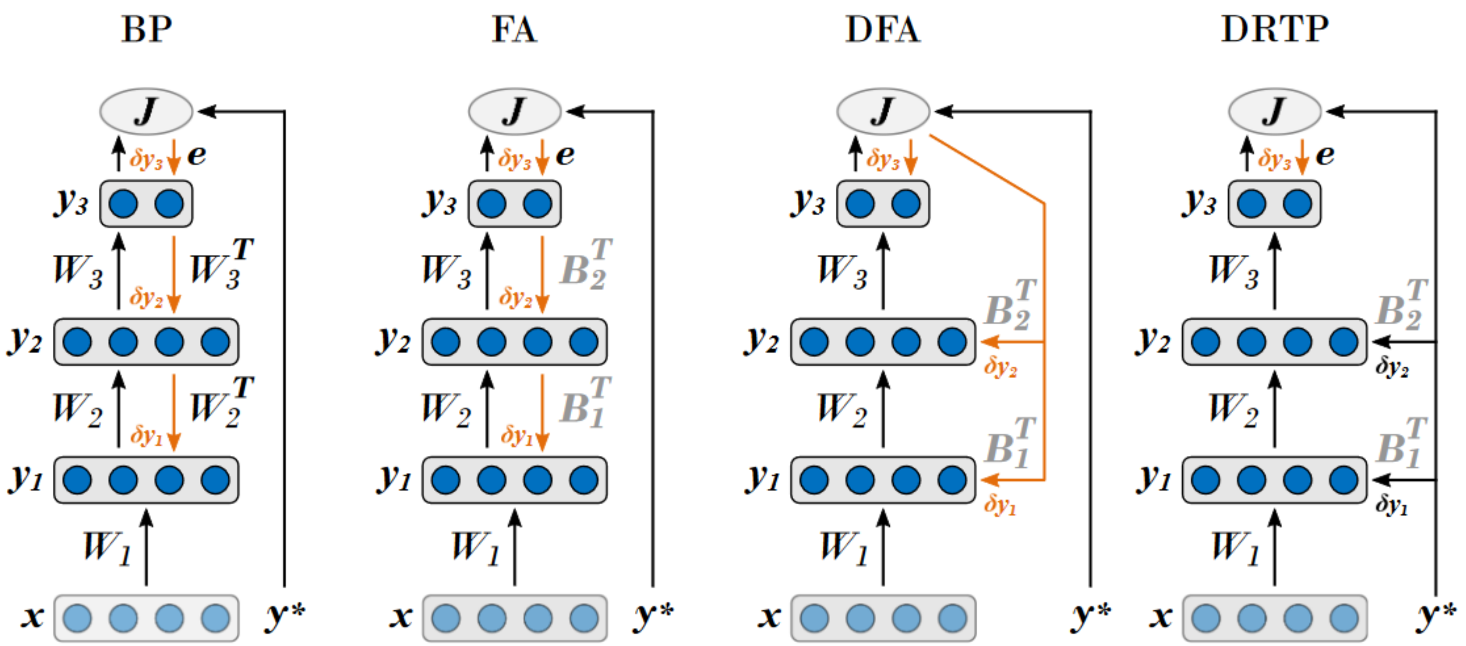
ICTEAM, Université catholique de Louvain, Belgium

‡ Equal contributions
*Now with Institute of Neuroinformatics, UZH and ETH Zürich, Switzerland
charlotte@ini.uzh.ch, martin.lefebvre@uclouvain.be

LightOn AI Meetup
Online, September 17, 2020

Overview – From feedback alignment to direct random target projection

Releasing the weight transport and update locking of backprop



δy_k	$\frac{\partial J}{\partial y_k} = W_{k+1}^T \delta z_{k+1}$	$B_k^T \delta z_{k+1}$	$B_k^T e$	$B_k^T y^*$
Weight-transport-free	×	✓	✓	✓
Update-unlocked	×	×	×	✓

Feedforward local training

↘ Computational and memory cost ↘

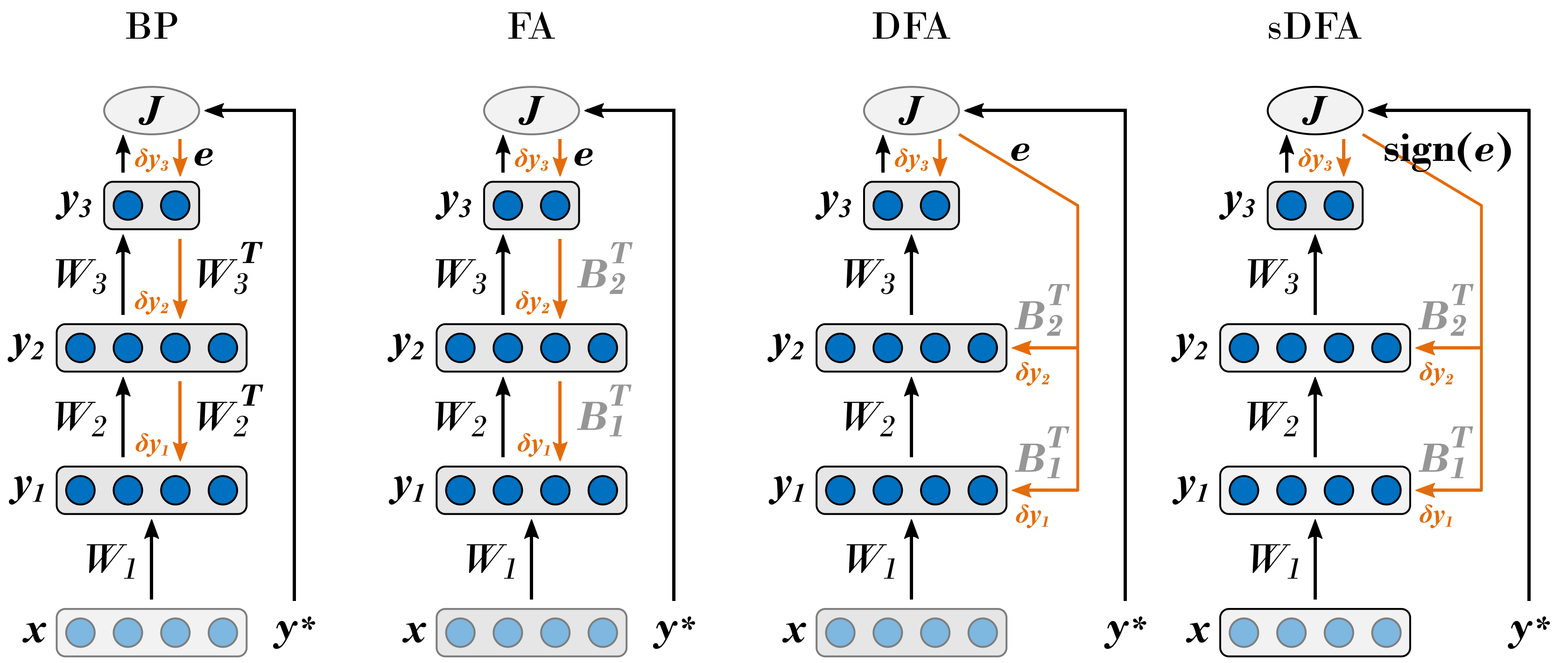
Outline

- Sign-based DFA (sDFA) solves synthetic regression and classification tasks
- From sDFA to DRTP: releasing update locking for classification tasks
- DRTP solves classification tasks: MNIST and CIFAR-10 benchmarking
- Conclusion and perspectives

Empirical results on synthetic datasets

Training algorithms

Claim: Weight updates based only on the error sign provide learning to multi-layer networks

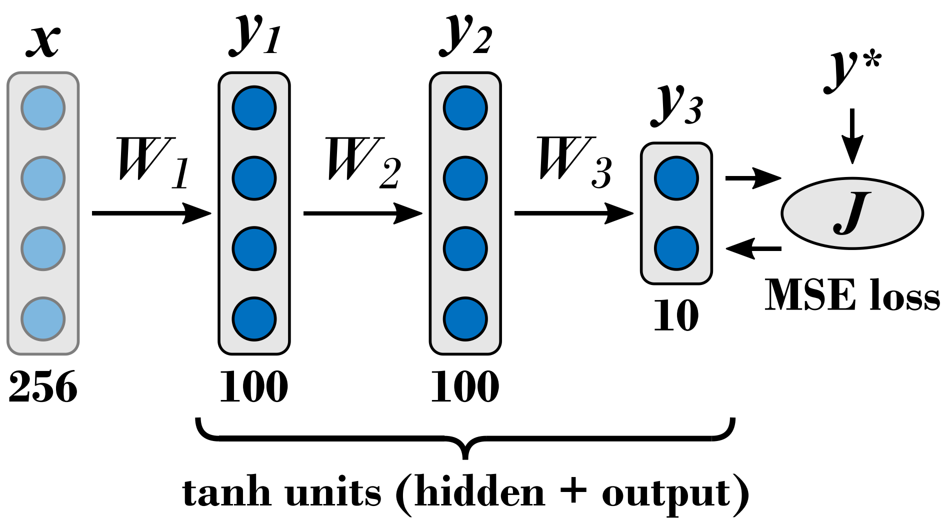


δy_k $\quad \frac{\partial J}{\partial y_k} = W_{k+1}^T \delta z_{k+1}$ $\quad B_k^T \delta z_{k+1}$ $\quad B_k^T e$ $\quad B_k^T \mathrm{sign}(e)$

Sign-based DFA solves regression tasks

Setup

Goal: Approximate 10 non-linear cosine functions



$y_i^* = \cos(\bar{x} + \phi_i)$ with $i \in \mathbb{Z}, i \in [0; 9]$

where $\phi_i = \frac{-\pi}{2} + \frac{i\pi}{9}$

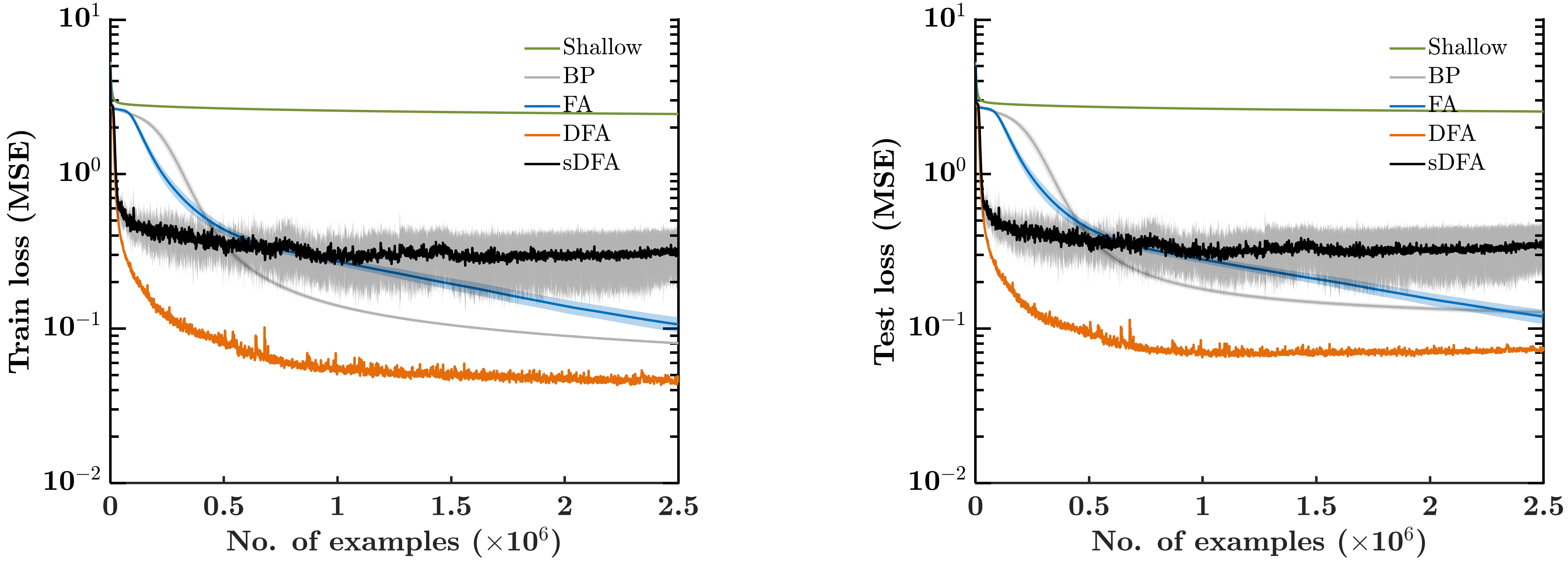
$\bar{x} = \text{mean}(x)$

$x_j \sim N(\mu_x, 1)$ with $j \in \mathbb{Z}, j \in [0; 255]$

$\mu_x \sim U(-\pi, \pi)$

Sign-based DFA solves regression tasks

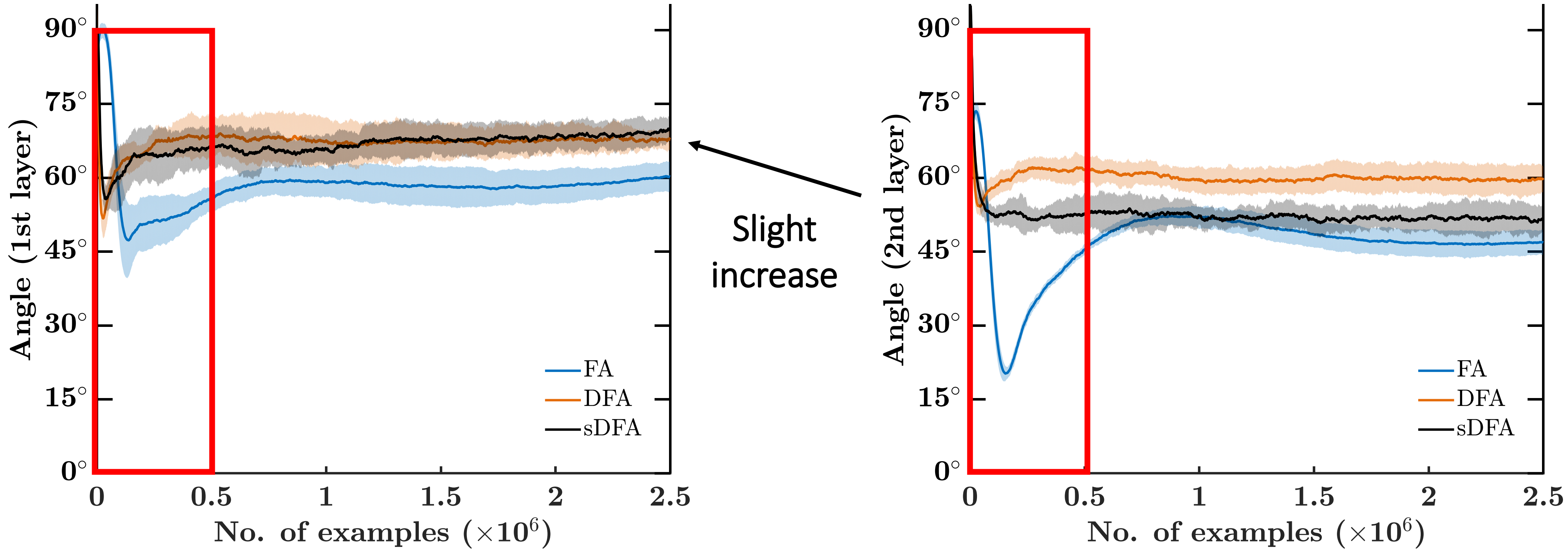
Loss



- DFA > BP/FA > sDFA > shallow
- sDFA performance drop due to lack of error magnitude information
  - No reduction of effective learning rate as training progresses
  - Class-dependent error magnitude is lost

Sign-based DFA solves regression tasks

Angle

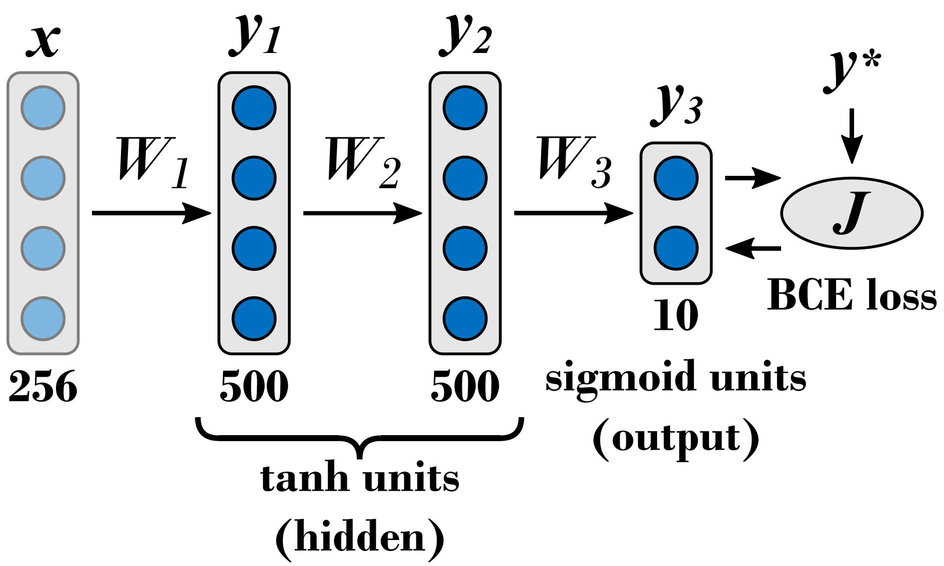


- FA alignment better during 100 first epochs
- sDFA alignment similar to DFA
- Alignment slightly **degrades** away from the network output

Sign-based DFA solves classification tasks

Setup

Goal: Classify 256-dimensional vectors into 10 classes

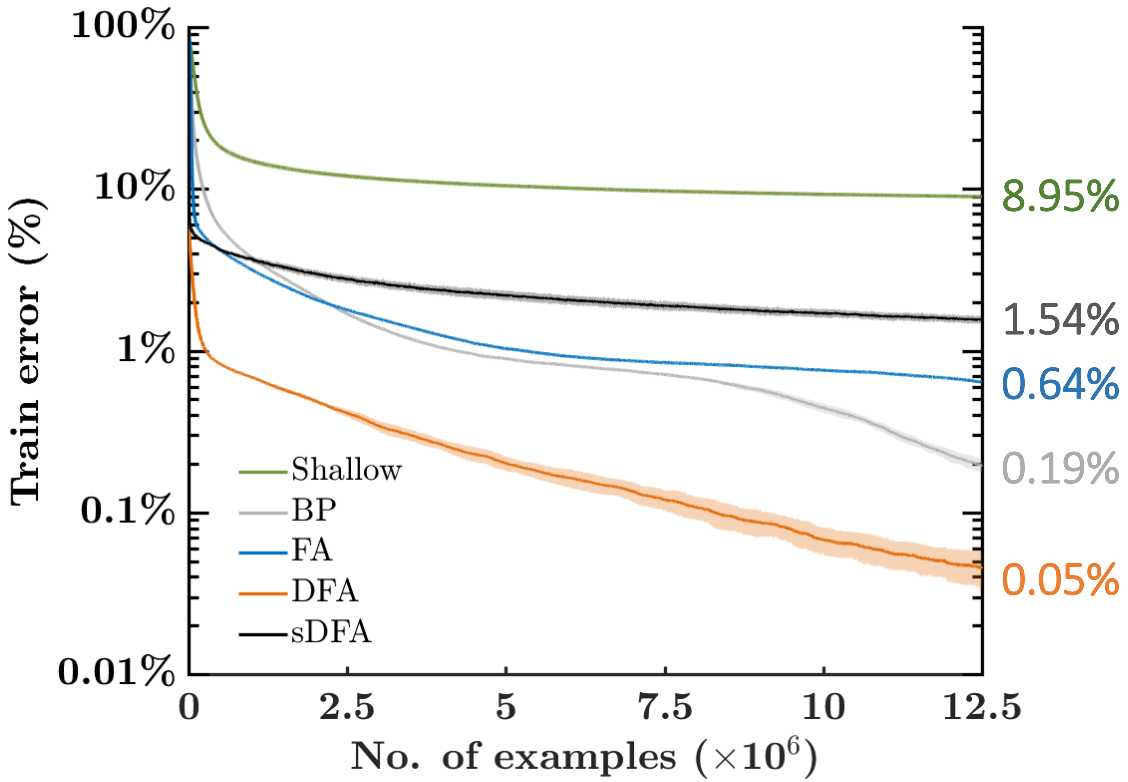


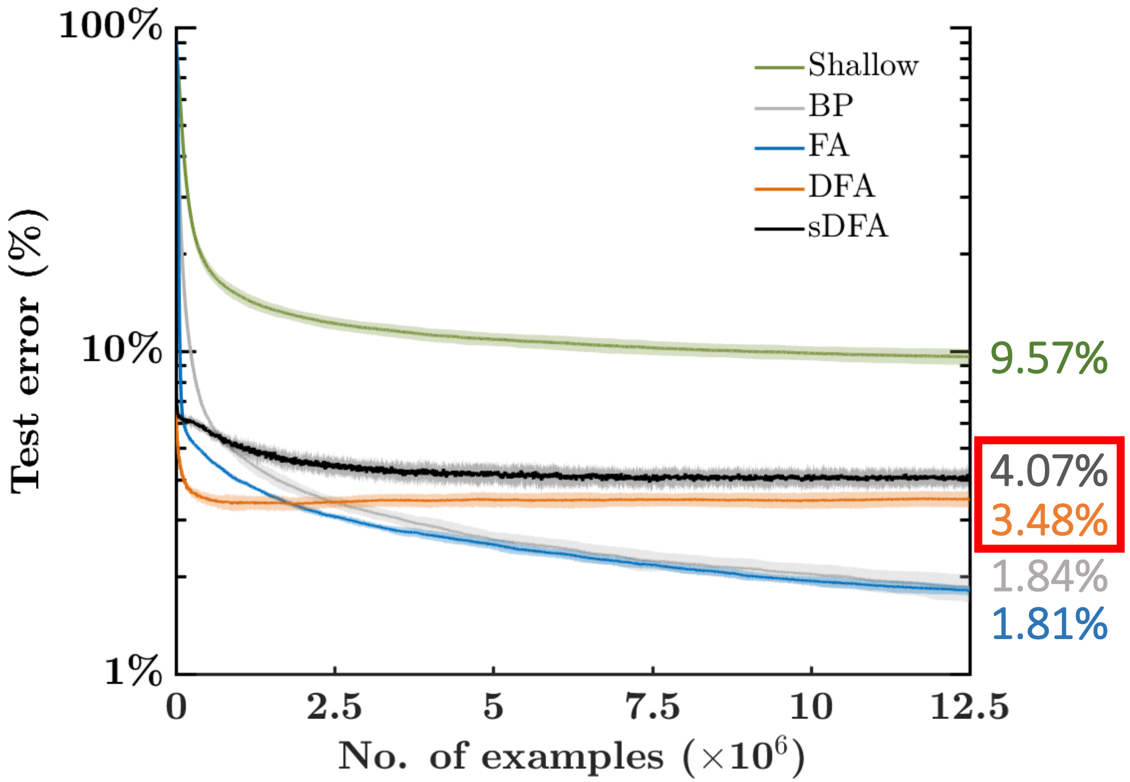
(x, y^*) pairs generated by sklearn library

(make_classification function)

Sign-based DFA solves classification tasks

Accuracy

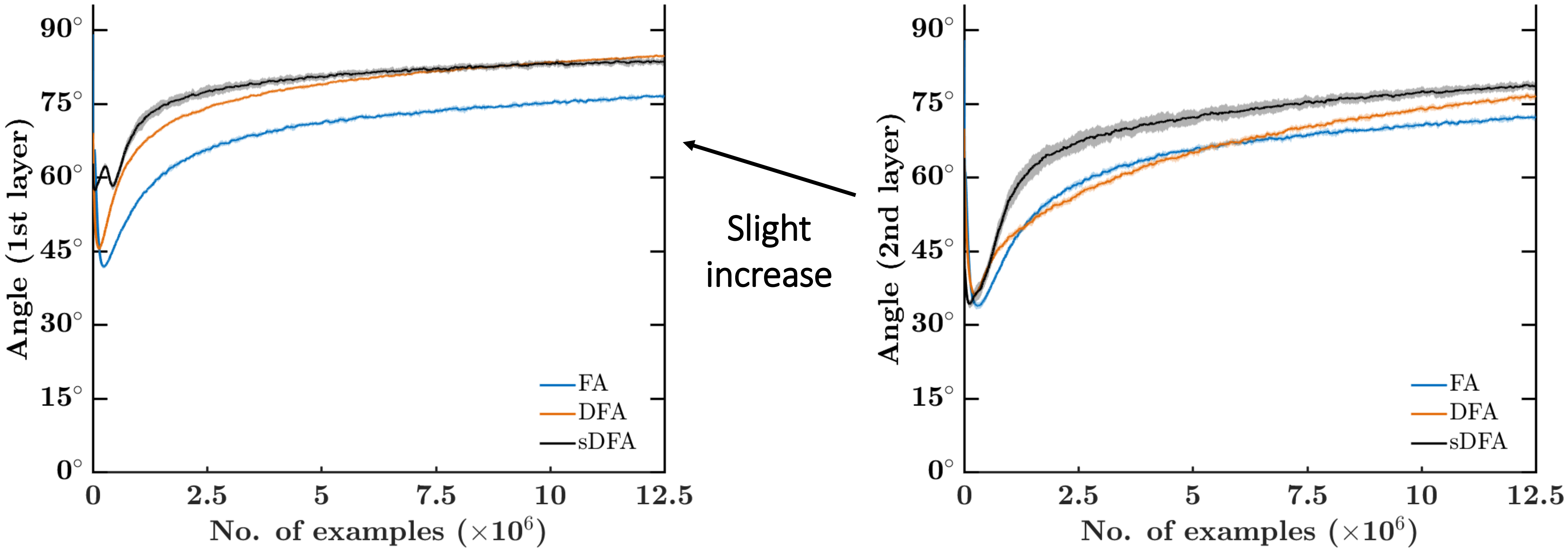




- **Training set**: DFA > BP > FA > sDFA > shallow
- **Test set**: BP/FA > DFA/sDFA > shallow

Sign-based DFA solves classification tasks

Angle



- FA > DFA/sDFA

Outline

- Sign-based DFA (sDFA) solves synthetic regression and classification tasks
- From sDFA to DRTP: releasing update locking for classification tasks
- DRTP solves classification tasks: MNIST and CIFAR-10 benchmarking
- Conclusion and perspectives

The error sign is known in advance

Claim: For classification, a feedback pathway is no longer needed as the error sign is known in advance

$$e = y^* - y_K \longrightarrow e_c = \begin{cases} 1 - y_{Kc} & \text{if} \quad c = c^* \\ -y_{Kc} & \text{otherwise} \end{cases}$$

Correct class
Incorrect classes

$y_{Kc} \in [0; 1]$ softmax/sigmoid

$$\downarrow$$

$$\text{sign}(e_c) = \begin{cases} +1 & \text{if} \quad c = c^* \\ -1 & \text{otherwise} \end{cases}$$

Correct class
Incorrect classes

For a given example, the error sign does not change during training

DRTP solves classification tasks without feedback

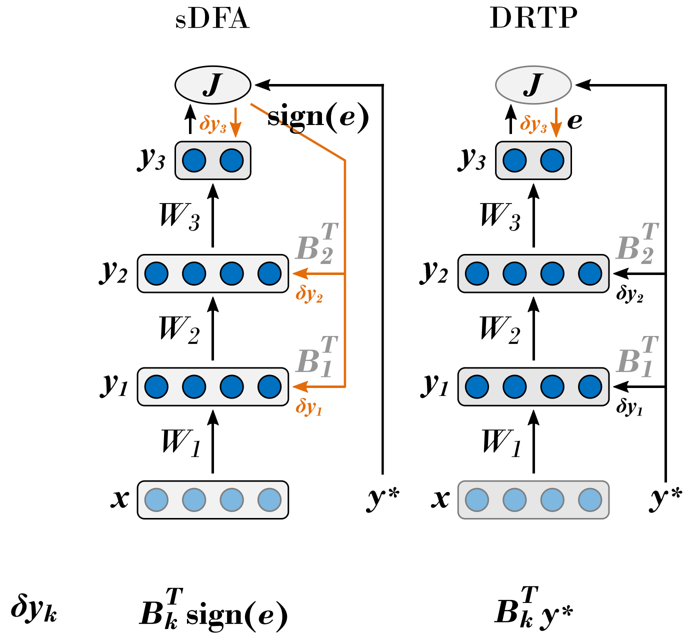
Link between sDFA and DRTP

Claim: Direct random target projection (DRTP) delivers useful modulatory signals

DRTP is a simplified version of sDFA

1. DRTP is computationally cheaper than sDFA
   DRTP: Label-dependent selection of layerwise random vector
   sDFA: Matrix product between error vector and fixed random matrix

2. DRTP systemically outperforms sDFA on MNIST and CIFAR-10 datasets
   DRTP: Only the correct class has an impact
   sDFA: The C-1 incorrect classes outweigh the correct class



The directions of DRTP and BP modulatory signals are within 90° of each other

➡ See full proof in the paper

Outline

- Sign-based DFA (sDFA) solves synthetic regression and classification tasks
- From sDFA to DRTP: releasing update locking for classification tasks
- DRTP solves classification tasks: MNIST and CIFAR-10 benchmarking
- Conclusion and perspectives

DRTP benchmarking on classification tasks

MNIST assessment

MNIST

Network	BP	FA	DFA	DRTP
784-500-10	1.72±0.08%	1.92±0.08%	2.59±0.11%	4.58±0.12%
784-1000-10	1.76±0.06%	1.90±0.06%	2.12±0.05%	4.03±0.13%
784-500-500-10	1.62±0.12%	1.95±0.07%	4.35±0.30%	4.57±0.13%
784-1000-1000-10	1.67±0.07%	1.90±0.07%	3.46±0.25%	4.04±0.12%
CONV* (random)	1.31±0.08%	1.55±0.04%	1.66±0.11%	1.87±0.12%
CONV* (trained)	0.99±0.05%	1.38±0.06%	2.38±0.39%	1.81±0.14%

* 28x28-32c5-2p-1000-10

Fully-connected networks

Performance degrades as the BP constraints are relaxed

DRTP is still competitive

DRTP benchmarking on classification tasks

MNIST assessment

MNIST

Network	BP	FA	DFA	DRTP
784-500-10	1.72±0.08%	1.92±0.08%	2.59±0.11%	4.58±0.12%
784-1000-10	1.76±0.06%	1.90±0.06%	2.12±0.05%	4.03±0.13%
784-500-500-10	1.62±0.12%	1.95±0.07%	4.35±0.30%	4.57±0.13%
784-1000-1000-10	1.67±0.07%	1.90±0.07%	3.46±0.25%	4.04±0.12%
CONV* (random)	1.31±0.08%	1.55±0.04%	1.66±0.11%	1.87±0.12%
CONV* (trained)	0.99±0.05%	1.38±0.06%	2.38±0.39%	1.81±0.14%

* 28x28-32c5-2p-1000-10

Convolutional neural networks (fixed random convolutional layers)

All training algorithms lie close to each other on MNIST

DRTP benchmarking on classification tasks

MNIST assessment

MNIST

Network	BP	FA	DFA	DRTP
784-500-10	1.72±0.08%	1.92±0.08%	2.59±0.11%	4.58±0.12%
784-1000-10	1.76±0.06%	1.90±0.06%	2.12±0.05%	4.03±0.13%
784-500-500-10	1.62±0.12%	1.95±0.07%	4.35±0.30%	4.57±0.13%
784-1000-1000-10	1.67±0.07%	1.90±0.07%	3.46±0.25%	4.04±0.12%
CONV* (random)	1.31±0.08%	1.55±0.04%	1.66±0.11%	1.87±0.12%
CONV* (trained)	0.99±0.05%	1.38±0.06%	2.38±0.39%	1.81±0.14%

* 28x28-32c5-2p-1000-10

Convolutional neural networks (trained convolutional layers)

Only BP allows leveraging training for convolutional layers.
Feedback-alignment-based algorithms require parameter redundancy, which is not offered in convolutional layers.

DRTP benchmarking on classification tasks

MNIST and CIFAR-10 assessment

MNIST

Network	BP	FA	DFA	DRTP
784-500-10	1.72±0.08%	1.92±0.08%	2.59±0.11%	4.58±0.12%
784-1000-10	1.76±0.06%	1.90±0.06%	2.12±0.05%	4.03±0.13%
784-500-500-10	1.62±0.12%	1.95±0.07%	4.35±0.30%	4.57±0.13%
784-1000-1000-10	1.67±0.07%	1.90±0.07%	3.46±0.25%	4.04±0.12%
CONV* (random)	1.31±0.08%	1.55±0.04%	1.66±0.11%	1.87±0.12%
CONV* (trained)	0.99±0.05%	1.38±0.06%	2.38±0.39%	1.81±0.14%

* 28x28-32c5-2p-1000-10

CIFAR-10

Network	BP	FA	DFA	DRTP
784-500-10	48.43±0.30%	49.59±0.25%	49.73±0.24%	53.72±0.30%
784-1000-10	47.58±0.21%	48.56±0.28%	48.45±0.17%	52.99±0.22%
784-500-500-10	49.23±0.24%	50.83±0.20%	50.76±0.24%	53.46±0.16%
784-1000-1000-10	49.00±0.22%	50.35±0.18%	50.51±0.24%	52.83±0.44%
CONV* (random)	30.13±0.31%	30.28±0.37%	30.40±0.46%	32.69±0.38%
CONV* (trained)	27.45±0.28%	29.84±0.31%	32.06±0.29%	35.45±0.76%

* 32x32x3-64c3-2p-256c3-2p-1000-1000-10

Outline

- Sign-based DFA (sDFA) solves synthetic regression and classification tasks
- From sDFA to DRTP: releasing update locking for classification tasks
- DRTP solves classification tasks: MNIST and CIFAR-10 benchmarking
- Conclusion and perspectives

Take-home messages

1. The error sign is sufficient to provide learning to multi-layer networks

2. For classification problems, the error sign is known in advance
   Solves the update locking problem!

3. ‘Soft’ alignment between forward and backward weights
   $\Leftrightarrow$ Modulatory signals within 90° of those prescribed by BP

4. DRTP is demonstrated on the MNIST and CIFAR-10 datasets

Outlook

Neuroscience

DRTP could come in line with recent findings in cortical areas that reveal the existence of output-independent target signals in the dendritic instructive pathways of intermediate-layer neurons.

[Magee & Grienberger, *Annual Review of Neuroscience*, 2020]

Circuit implementation

Can lead to record low silicon area and energy overheads to embed on-chip online learning for edge computing devices.

[Frenkel, ISCAS, 2020]

Thank you!

Further resources:

The DRTP preprint: https://arxiv.org/pdf/1909.01311.pdf

Open-source DRTP PyTorch code: https://github.com/chfrenkel

ISCAS paper for the silicon implementation: https://arxiv.org/pdf/2005.06318.pdf

Supplementary – Synthetic datasets

Learning parameters

Regression task: 5k training set, 1k test set
Classification task: 25k training set, 5k test set

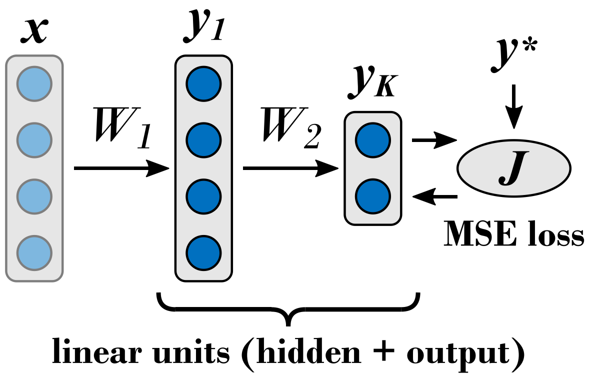
- 500 epochs
- Mini-batches of size 50
- Learning rate of 5×10^{-4}, similar for all training algorithms
- Forward weights (W_k) are drawn from He distributions (BP) or zero-initialized for FA-based training algorithms
- Feedback weights (B_k) are drawn from He distributions

Supplementary - Classification problems

DRTP proof of alignment

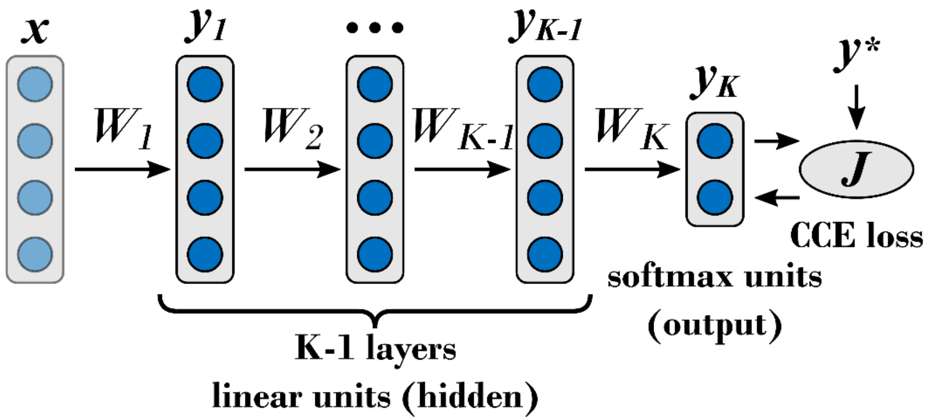
Claim: Direct random target projection delivers useful modulatory signals
The directions of DRTP and BP modulatory signals are within 90° of each other

FA proof [Lillicrap, *Nat. Comms.*, 2016]



- Single example
- Forward weights zero-initialized
- 1 linear hidden layer
- Linear output layer
- MSE loss

DRTP proof



- Single example
- Forward weights zero-initialized
- K-1 (arbitrary no.) linear hidden layers
- Softmax/sigmoid output layer
- CCE/BCE loss

Supplementary - Classification problems

DRTP proof of alignment

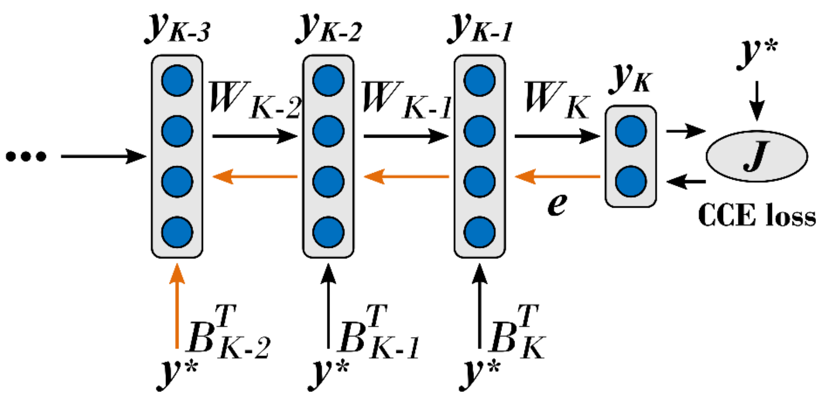
Claim: Direct random target projection delivers useful modulatory signals
The directions of DRTP and BP modulatory signals are within 90° of each other

BP modulatory signals

$$\delta y_k = \delta z_k = -\frac{1}{C}\left(\prod_{i=k+1}^{K} W_i^T\right) e$$

DRTP modulatory signals

$$\delta y_k = \delta z_k = B_k^T y^*$$



Alignment

$$-\frac{1}{C} e^T \left(\prod_{i=k+1}^{K} W_i^T\right)^T B_k^T y^* > 0$$

Theorem

$$B_k^T y^* = -\alpha_k^t \left(\prod_{i=K}^{k+1} W_i^T\right)^+ e$$

where $\alpha_k^t > 0$