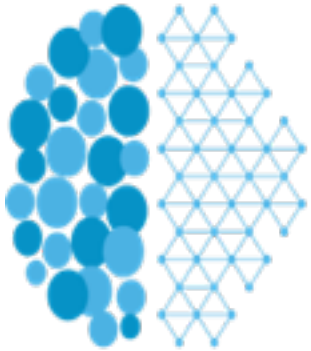
More Steps Towards Biologically Plausible Backprop

Yoshua Bengio

December 9, BigNeuro NIPS 2017 Workshop

Long Beach, USA



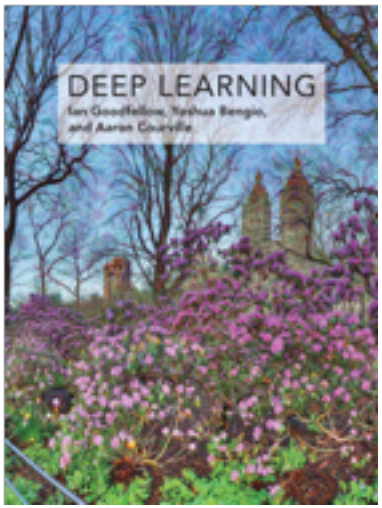
MILA

CIFAR
CANADIAN INSTITUTE FOR ADVANCED RESEARCH

Université de Montréal

IVADO

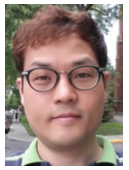
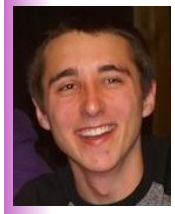
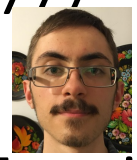
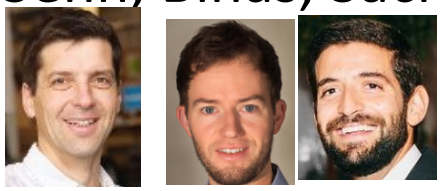
PLUG: **Deep Learning**, MIT Press book is out, chapters will remain online



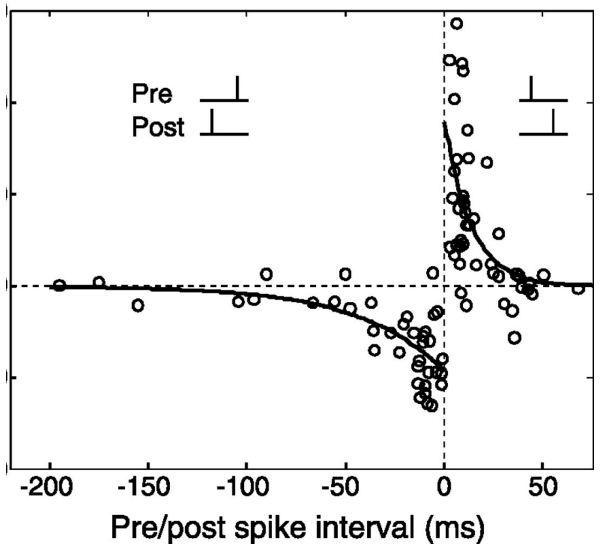
Deep Learning & Neuroscience: Still a Large Gap

- **Backprop** and the ability to jointly train multiple layers is the workhorse of current deep learning successes. END-TO-END TRAINING OF DEEP COMPUTATIONS ROCKS. **Backprop is the building block behind modern unsupervised (generative) learning and RL**. But has been deemed not biologically plausible.
  - how to propagate gradients? linear neurons? separate net?
  - what is the role of feedback connections? lateral connections?
  - How to **efficiently** train a stochastic continuous-time dynamical system wrt a **global** objective?
    - *Random perturbation-based methods do not scale, BP does beautifully*

Towards Key Principles of Backprop-Like Learning in Brains: Neural Nets → Brain implementation

- Joint training of many areas in the brain is still a mystery
- We have made some progress in bridging the gap between backprop and the brain
  - Lee et al (Difference TargetProp) ECML'2015 
  - Bengio et al (STDP – CHL) Neural Computation, 2017  
  - Bengio & Fischer (Neural inference) arXiv 1510.02777
  - Bengio et al (feedforward init arXiv:1606.01651) 
  - Scellier & Bengio arXiv (Equilibrium Propagation) Frontiers in Neurosc. 2017 
  - Senn, Binas, Sacramento & Bengio – in progress (mirror interneurons for linearized feedback, avoids to wait for convergence to update W) 

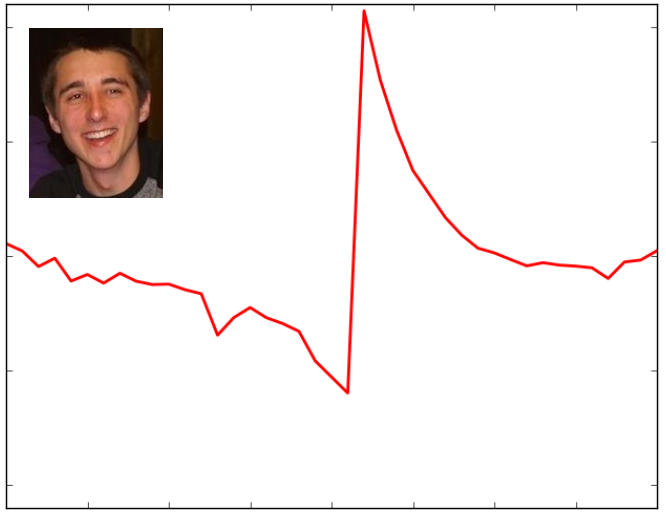
The STDP Connection



- Inspiration from Hinton 2007 (talk at Deep Learning Workshop @ NISP); see also April 2016 talk by Hinton @ Stanford, "Can the brain do back-propagation?"
- *Bengio et al 2015 & 2017 "STDP-compatible approximation of backpropagation in an energy-based model" arXiv:1509.05936, Neural Computation 2017*
  - shows that weight updates

$$\Delta W_{i,j} \propto \frac{d\rho(s_i)}{dt}\rho(s_j)$$

  - replicates the STDP experimental signature. If symmetry is added we get the same weight update as Eq.Prop.



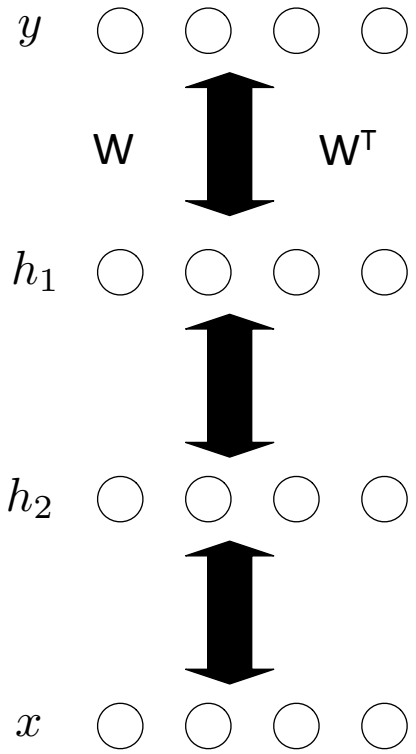
Nudge on output units propagates like backpropagated gradients

Bengio & Fischer, 2015, arXiv:1510.02777

Variation on the output *y* is propagated into a variation in h_1 mediated by the feedback weights W^T = transpose of feedforward weights W

Then the variation in h_1 is transformed into a variation in h_2, etc.

And we show that $\dot{h}$ proportional to direction of descent for the cost C



From Deep Learning to Neuroscience
Propagation of Error Signals

Deep Learning

Backpropagation:
Requires a special computational path for the propagation of error derivatives backward in the network.

Neuroscience

Hypothesis: ↔ *no empirical evidence yet*
Error signals are encoded in ds/dt.
No need for a special computational path.

This idea was first proposed by Hinton & McClelland:
"Recirculation algorithm" (1987)

Equilibrium Propagation
Continuous Hopfield Model Revisited

- Energy Function

$$E := \frac{1}{2}\left(\|h_2\|^2 + \|h_1\|^2 + \|h_0\|^2\right) - \left(\rho(x)^T W_{32}\rho(h_2) + \rho(h_2)^T W_{21}\rho(h_1) + \rho(h_1)^T W_{10}\rho(h_0)\right)$$

- Cost Function

$$C := \frac{1}{2}\|h_0 - \mathrm{y}\|^2$$

target output

- Objective Function

(= Cost at the Fixed Point)

$$J := C\left(s^0\right)$$

fixed point (local minimum of E)

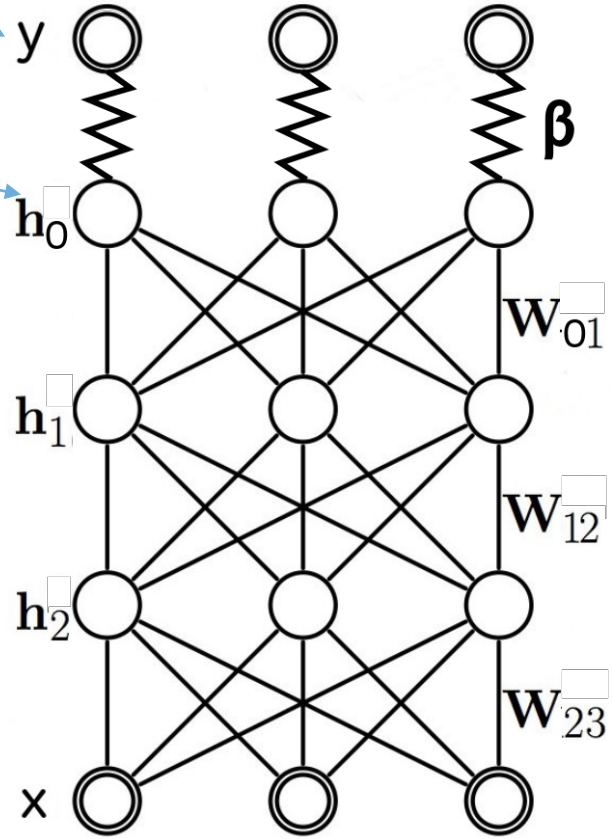
- **Extended Energy Function**

$$F := \underbrace{E}_{\text{'internal potential'}} + \underbrace{\beta C}_{\text{'external potential'}}$$

clamping factor **β≥0**

models interactions in the network

attracts h_0 towards y if **β>0**



output (= prediction)

input clamped

$$s = (h_0, h_1, h_2)$$

Equilibrium Propagation
Continuous Hopfield Model Revisited

- Neuronal Dynamics

$$\frac{ds}{dt} = -\frac{\partial F}{\partial s} = \underbrace{-\frac{\partial E}{\partial s}}_{\text{leaky integration}} \quad \underbrace{-\beta\frac{\partial C}{\partial s}}_{\text{'external influence'}}$$

follow the gradient of the Extended Energy

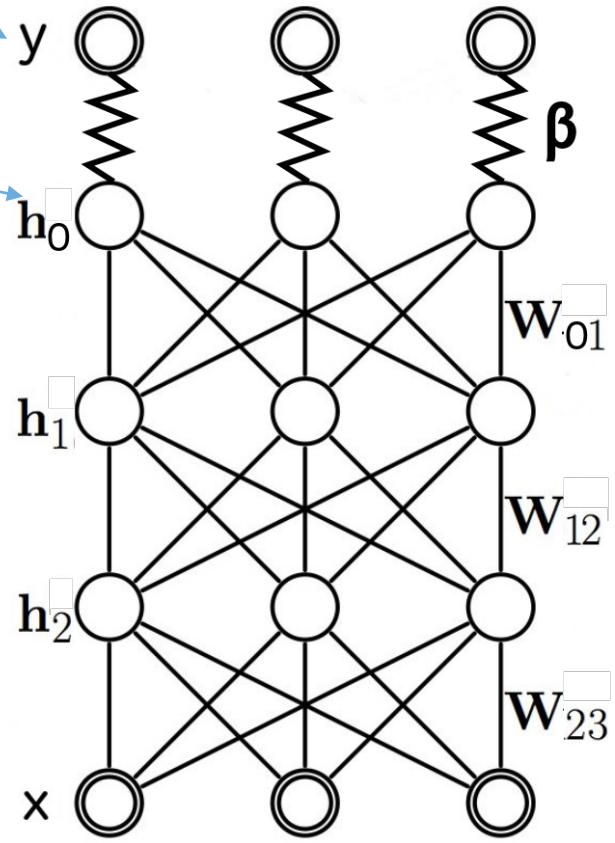
$$-\beta\frac{\partial C}{\partial h_0} = \beta(y - h_0)$$

pushes h_0 towards y if **β>0**

$$-\beta\frac{\partial C}{\partial h_i} = 0$$

target output ↔ y

output (= prediction) ↔ h_0



input clamped ↔ x

$$s = (h_0, h_1, h_2)$$

- **Extended Energy Function**

$$F := \underbrace{E}_{\text{'internal potential'}} + \underbrace{\beta C}_{\text{'external potential'}}$$

clamping factor **β≥0**

models interactions in the network

attracts h_0 towards y if **β>0**

Equilibrium Propagation
Continuous Hopfield Model Revisited

First Phase

(**Free** Phase)

- clamped inputs
- **free** outputs (**β=0**)
- seek for a minimum of E

Dynamics $$\frac{ds}{dt} = -\frac{\partial E}{\partial s}$$

Free Fixed Point

$$s^0 \quad \text{s.t.} \quad \frac{\partial E}{\partial s}\left(s^0\right) = 0$$

Second Phase

(**Weakly Clamped** Phase)

- clamped inputs
- **weakly clamped** outputs (**β≳0**)
- seek for a minimum of $E + \beta C$

Dynamics $$\frac{ds}{dt} = -\frac{\partial E}{\partial s} - \beta\frac{\partial C}{\partial s}$$

Weakly Clamped Fixed Point

$$s^\beta \quad \text{s.t.} \quad \frac{\partial E}{\partial s}\left(s^\beta\right) + \beta\frac{\partial C}{\partial s}\left(s^\beta\right) = 0$$

has lower cost value than s⁰

$$\Delta W_{ij} \propto \lim_{\beta \to 0} \frac{1}{\beta}\left(\rho\left(s_i^\beta\right)\rho\left(s_j^\beta\right) - \rho\left(s_i^0\right)\rho\left(s_j^0\right)\right)$$

update rule

decrease energy

increase energy

Equilibrium Propagation
(Scellier & Bengio 2017, Frontiers in Neuroscience)

Backpropagation

Free Phase
-network relaxes to fixed point
-read prediction at the outputs

Forward Pass
-read prediction at the outputs

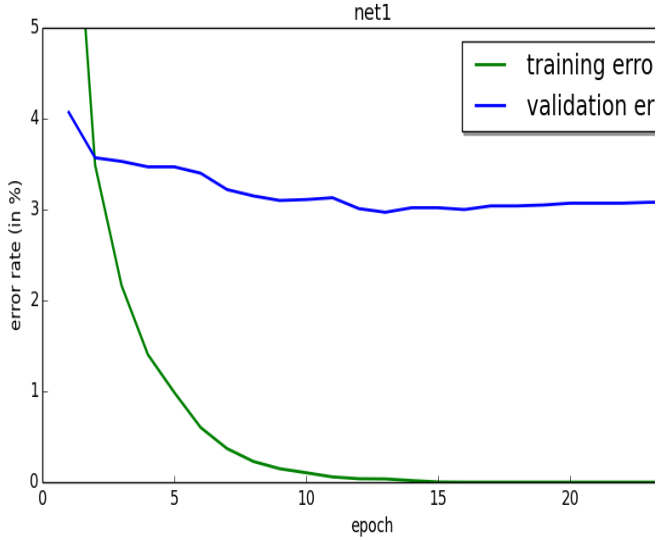
Weakly Clamped Phase
-nudge outputs towards targets
-error signals (back)propagate
-network relaxes to new nearby fixed point

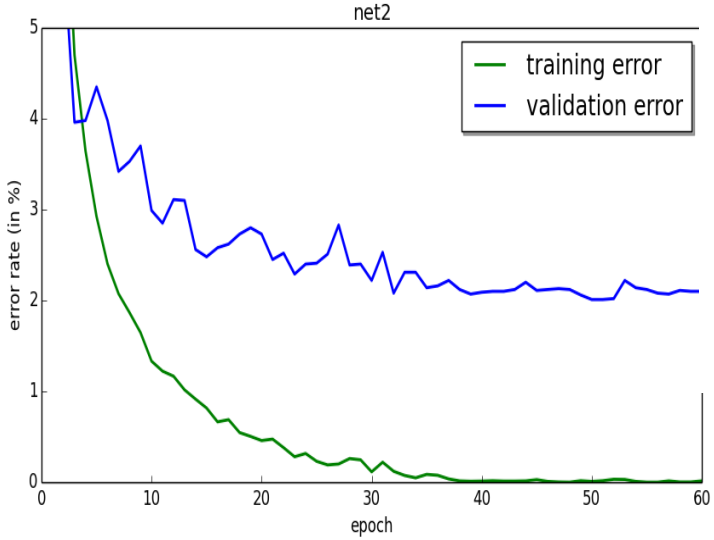
Backward Pass
-compare prediction/target
-compute error derivatives

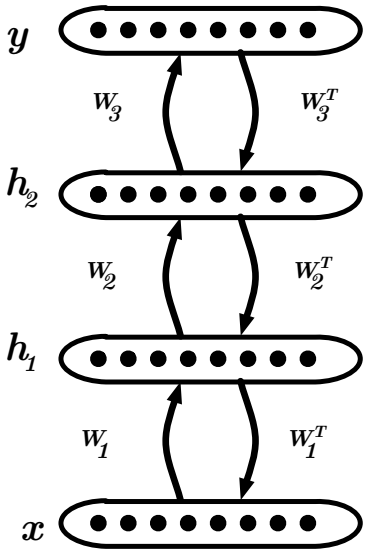
requires:
-special computational circuit
-special kind of computation

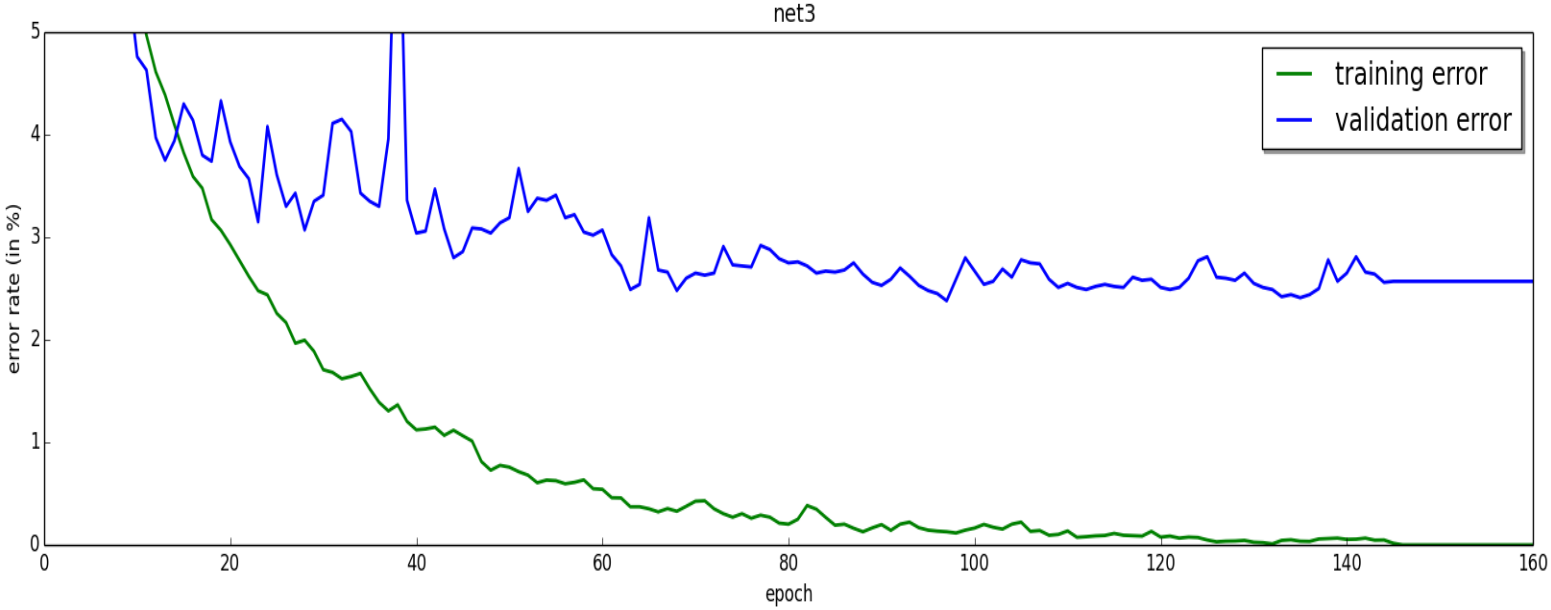
Simulation Results

Ok but only if we wait for lengthy convergence near f.p.









Implicit Feedback & Feedforward Energy Function

- The experiments lead to unrealistically long convergence time to fixed point, and an update which does not work well unless we are not close enough to the fixed point
- The feedback connections in the Hopfield energy function may not be necessary, because feedforward energy terms give rise to symmetric update terms

$$E = \frac{1}{2}\sum_k ||s_k - W_k\rho(s_{k-1})||^2$$

$$\frac{\partial E}{\partial s_k} = \underbrace{(s_k - W_k\rho(s_{k-1}))}_{\text{feedforward paths}} + \underbrace{W_{k+1}^T\rho'(s_k) \circ (s_{k+1} - W_{k+1}\rho(s_k))}_{\text{feedback paths}}$$

- Purely feedforward energy function → one-pass convergence

(see also Whittington & Bogacz, Neural Comp. 2017)

E=0 in prediction phase means that there is no need for two distinct phases

- The feedforward energy has another amazing side-effect, besides one-pass convergence at prediction time:
  - only one of the two phases of Equilibrium-Prop is necessary, the nudging phase, because the $\beta = 0$ term of the weight update vanishes

$$\frac{\partial E}{\partial W_k} = (s_k - W_k \rho(s_{k-1}))^T \rho'(s_{k-1}) \circ \rho(s_{k-1})$$

  - **= 0 at the the $\beta = 0$ fixed point** because it has

$$s_k = W_k \rho(s_{k-1})$$

  - The brain may be continuously 'chasing' a target, being slightly nudged towards it

Feedforward Energy Function→ BP

(see also Whittington & Bogacz, Neural Comp. 2017)

- The fixed point solution in the nudging phase $\beta > 0$ can also be done in a single (feedback) pass, which is exactly backprop:

$$\frac{\partial E}{\partial s_k} = \underbrace{(s_k - W_k \rho(s_{k-1}))}_{e_k} - W_{k+1}^T \rho'(s_k) \circ \underbrace{(s_{k+1} - W_{k+1}\rho(s_k))}_{e_{k+1}} + \beta \frac{\partial C}{\partial s_k}$$

Firing rate

nudging

- Setting $\frac{\partial E}{\partial s_k} = 0$, hidden layers should somehow compute

$$e_k = \underbrace{W_{k+1}^T \rho'(s_k) \circ}_{\text{Linear feedback pathway}} e_{k+1} \quad \text{where} \quad e_k = -\beta \frac{\partial C}{\partial \rho(s_k)}$$

- = backprop equations → but **linear neurons**, symmetric weights, which neurons compute *e*? isn't it biologically implausible?

Mirror Interneurons Solve the Linear Feedback Puzzle


with Walter Senn & Joao Sacramento

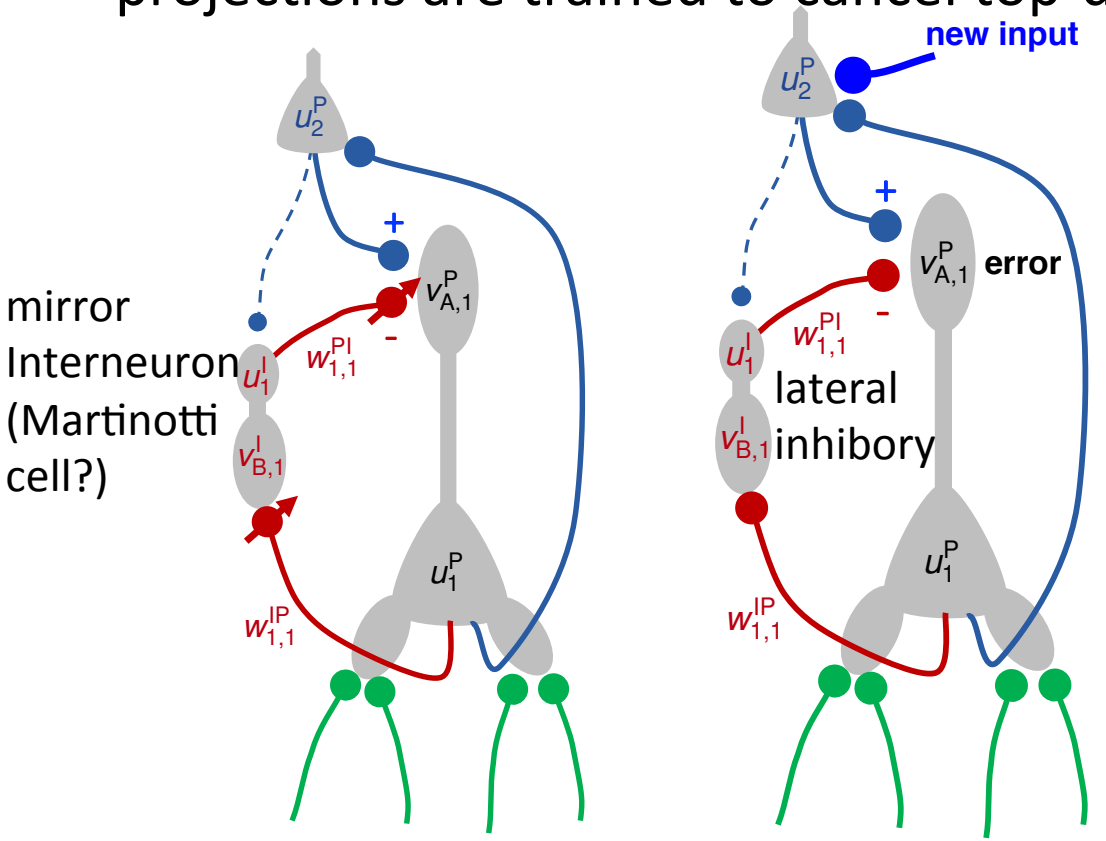
- Many attempts at biological implementations of backprop end up with the same problem: linear computation in backprop phase
- But neural computation is non-linear, and where would these errors be computed? Many unsatisfactory solutions have been proposed...
- NEW SOLUTION:
  - each pyramidal cell (main neuron, computes s_k) is associated with a mirror neuron g_k which imitates the feedforward component of s_k (obtained with no nudging), because g_k does not receive the top-down feedback weights
  - **error signal *e* = *bottom-up* – *top-down* inputs to main neuron**

Mirror Interneurons Solve the Linear Feedback Puzzle


Building on *Urbanczik & Senn 2014*

with Walter Senn & Joao Sacramento

- Mirror interneurons imitates feedforward path, their lateral projections are trained to cancel top-down feedback



Basal dendrites: bottom-up

Apical dendrites: top-down feedback minus mirror unit's cancellation.

With no nudging, cancellation is perfect because next layer is predictable.

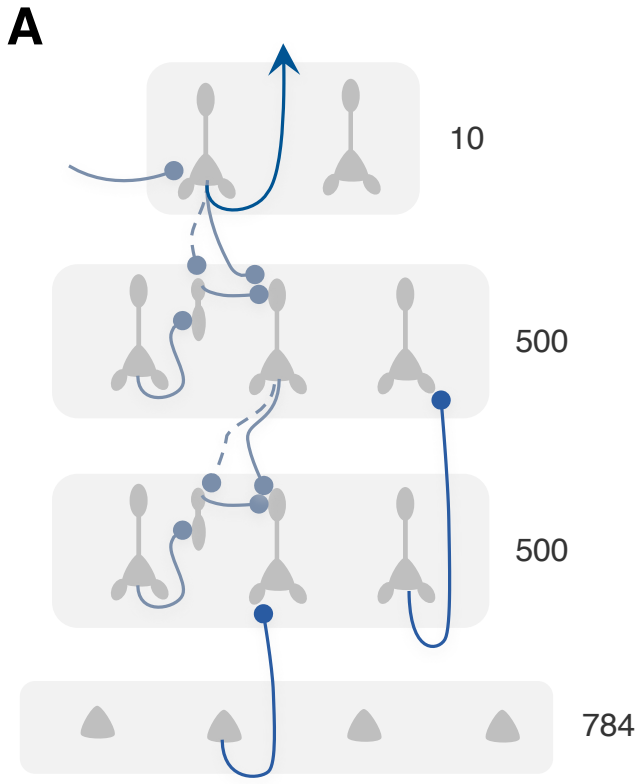
With nudging, difference = backprop error signal.

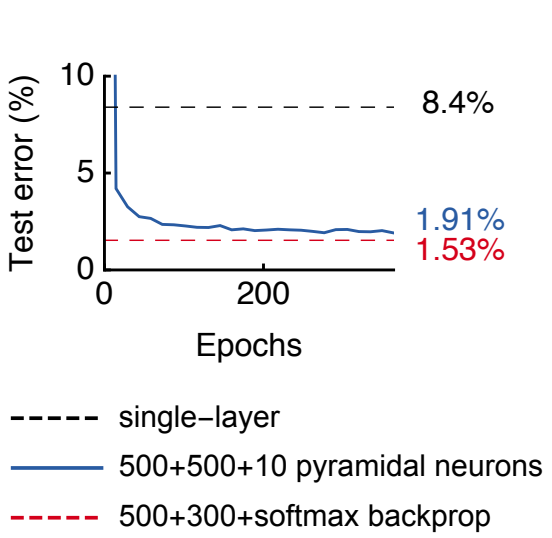
 See also *Dec. 5 Elife " Towards deep learning with segregated dendrites", Guerguiev et al*

MNIST Experiments with Mirror Interneurons


with Walter Senn & Joao Sacramento

Differential equations to simulate rate-based pyramidal and mirror neurons





Mirror Interneurons and Hinton's Temporal Derivative for Biological BP

- Theorem: if mirror paths are well trained (to cancel the feedback signals) and if feedback weights are symmetric and if the "errors" are small compared to the neuron's dynamic range, then

$$e_k \underset{\sim}{\propto} \frac{ds_k}{dt} \underset{\sim}{\propto} -\frac{\partial C}{\partial \rho(s_k)}$$

- Weight update $\Delta W_{ij} \propto e_i \rho'(s_i)\rho(s_j) \underset{\sim}{\propto} -\frac{\partial C}{\partial W_{ij}}$ follows the error gradient

Postsynaptic rate of change if active

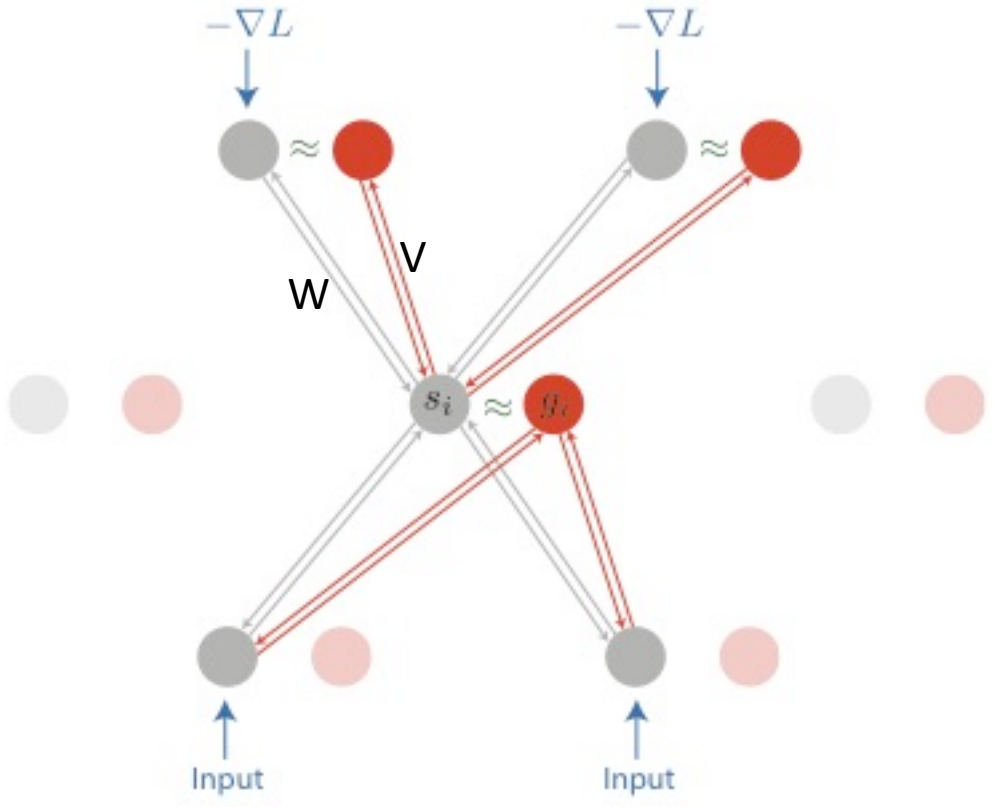
Presynaptic spikes

Simulation Results with Mirror Interneurons

Thomas Mesnard

Gaétan Vignoud

- Mirror units trained to match their corresponding pyramidal cell
- Continuous updates of V weights in prediction phase, of W weights in nudging phase

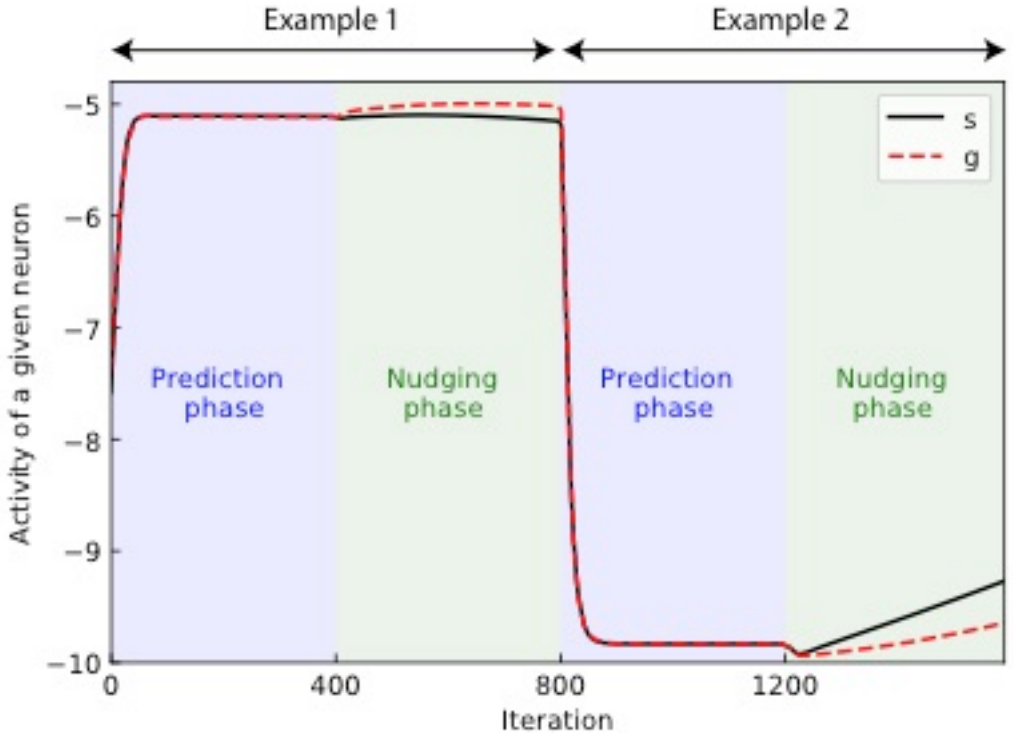


Simulation Results with Mirror Interneurons

Thomas Mesnard

Gaétan Vignoud

- During the nudging phase, the mirror units *g* do not match perfectly the feedback from the pyramidal units *s* downstream, leading to weight changes in W

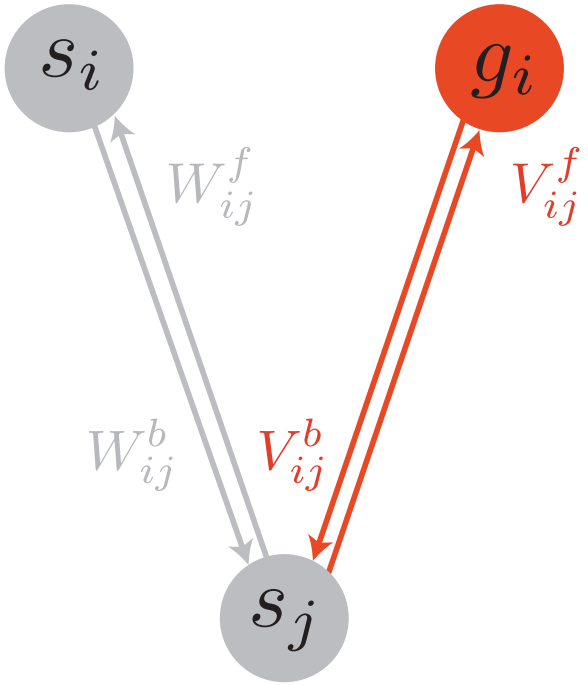


Simulation with Mirror Interneurons


Thomas Mesnard


Gaétan Vignoud

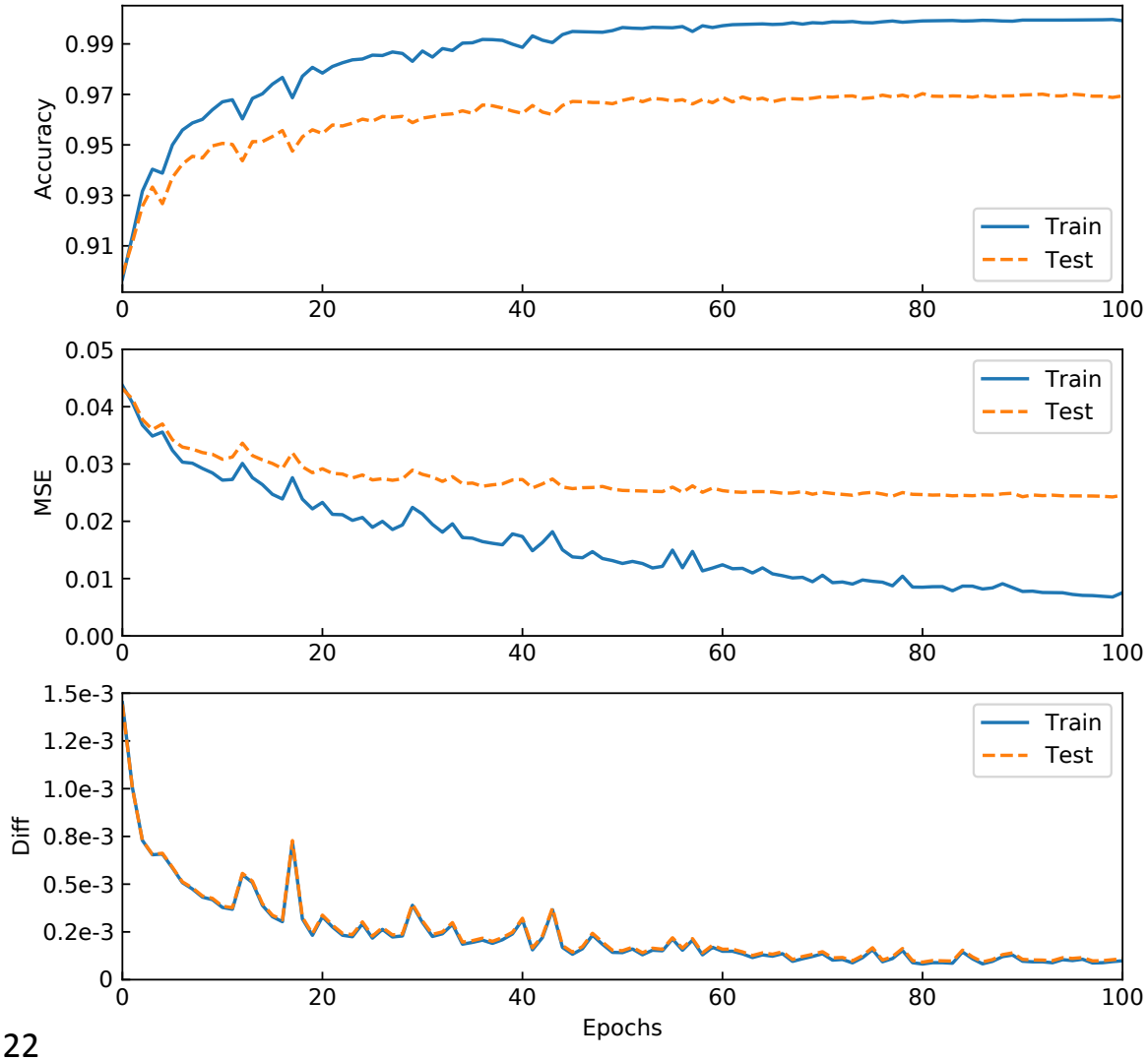
while *not done* **do**
  Sample batch from the training set
  for k *in range prediction_steps* **do**
    $$e_i = \left(\sum_{j \in S_i} W^b_{ij} \rho(s_j) - V^b_{ij} \rho(g_j), 0 \right)$$
    $$s_i \leftarrow s_i + \frac{\mathrm{d}t}{\tau} \left(-s_i + \sum_{j \in P_i} W^f_{ij} \rho(s_j) + e_i \right)$$
    $$g_i \leftarrow g_i + \frac{\mathrm{d}t}{\tau} \left(-g_i + \sum_{j \in P_i} V^f_{ij} \rho(s_j) \right)$$
    $$V^f_{ij} \leftarrow V^f_{ij} + \eta_V \, \mathrm{d}t \Big((s_i - g_i) \rho(s_j) \Big)$$
    $$V^b_{ij} \leftarrow V^b_{ij} + \eta_V \, \mathrm{d}t \Big(e_j \rho(g_i) \Big)$$
  end
  for k *in range nudging_steps* **do**
    $$e_i = \left(\sum_{j \in S_i} W^b_{ij} \rho(s_j) - V^b_{ij} \rho(g_j), -\beta \frac{\partial C}{\partial \rho(s)} \right)$$
    $$s_i \leftarrow s_i + \frac{\mathrm{d}t}{\tau} \left(-s_i + \sum_{j \in P_i} W^f_{ij} \rho(s_j) + e_i \right)$$
    $$g_i \leftarrow g_i + \frac{\mathrm{d}t}{\tau} \left(-g_i + \sum_{j \in P_i} V^f_{ij} \rho(s_j) \right)$$
    $$W^f_{ij} \leftarrow W^f_{ij} + \eta_W \, \mathrm{d}t \Big(e_i \rho'(s_i) \rho(s_j) \Big)$$
    $$W^b_{ij} \leftarrow W^f_{ji}$$
  end
end



Simulation Results with Mirror Interneurons


Thomas Mesnard


Gaétan Vignoud

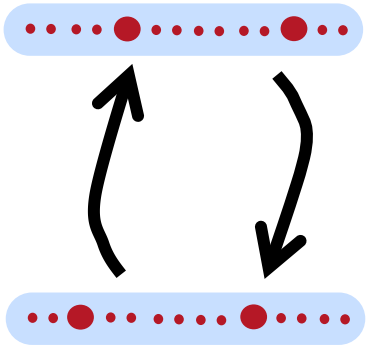


- MNIST results with 1 and 2 hidden layers
- Mirror interneurons track correctly

Weight Symmetry: maybe not a big deal

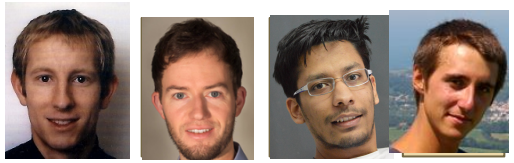
- The Feedback Alignment papers *(Lillicrap et al 2014)* show that weights used in backprop do not need to match perfectly the feedforward ones

- *Arora et al 2015* show that under sparsity and randomness assumptions, with rectifying non-linearity, the feedback weights which minimize one-layer reconstruction error are the symmetric weights



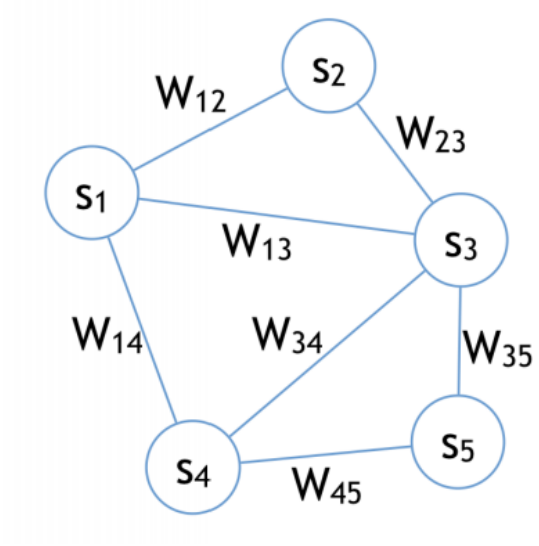
- Empirical results pre-confirmed this with denoising autoencoders, *Vincent et al 2010*.

Getting rid of symmetry requirements

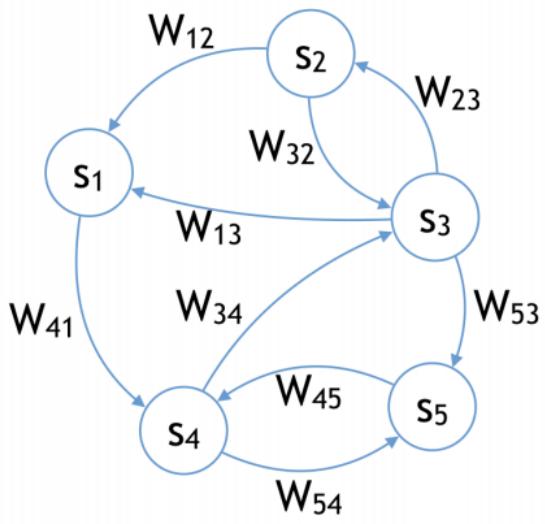


- Equilibrium propagation and related models are typically based on energy functions defined for systems with symmetric interactions, $w_{ij} = w_{ji}$.
- Such exact symmetry is unlikely to exist in biological systems.

- **Recent results:** equilibrium propagation can be extended to more general vector-field dynamics, without requiring an explicit energy function.



Energy-based model represented by an **undirected graph** (e.g. Hopfield net)



The generalized dynamics can be based on a **directed graph**.

Energy-based model

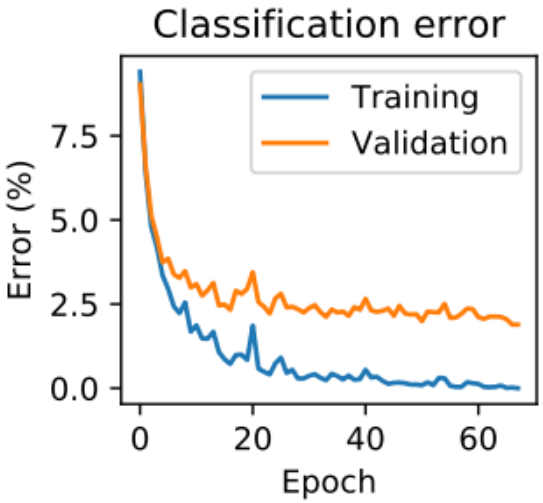
$$\frac{ds}{dt} = -\alpha \frac{\partial E}{\partial s}(W, s)$$

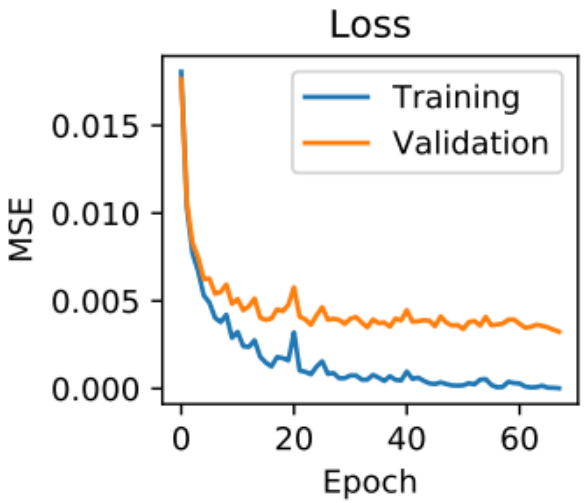
$$dW \propto \frac{\partial}{\partial W} \frac{\partial E}{\partial s} \cdot ds$$

Generalized vectorfield model

$$\frac{ds}{dt} = \mu(W, s)$$

$$dW \propto \frac{\partial \mu}{\partial W} \cdot ds$$





Experimental results

MNIST

784-512-512-512-10 MLP

200/100 iterations (1st/ 2nd phase)

- Initial experiments indicate that the generalized model learns faster than the energy-based one; this could be due to a greater number of free parameters (2x compared to a network with tied weights.)

Avoiding lengthy convergence: continuous update of the weights

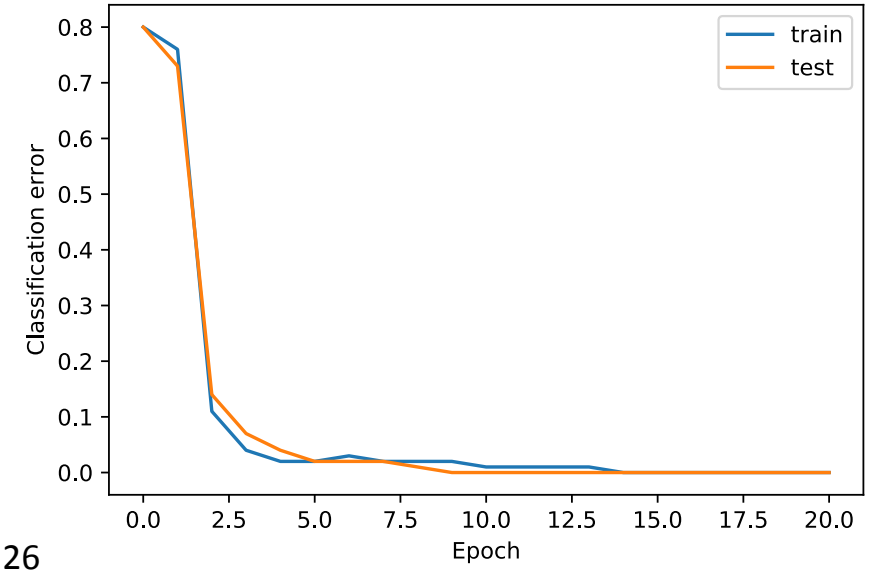
With W. Senn, J. Binas, J. Sacramento

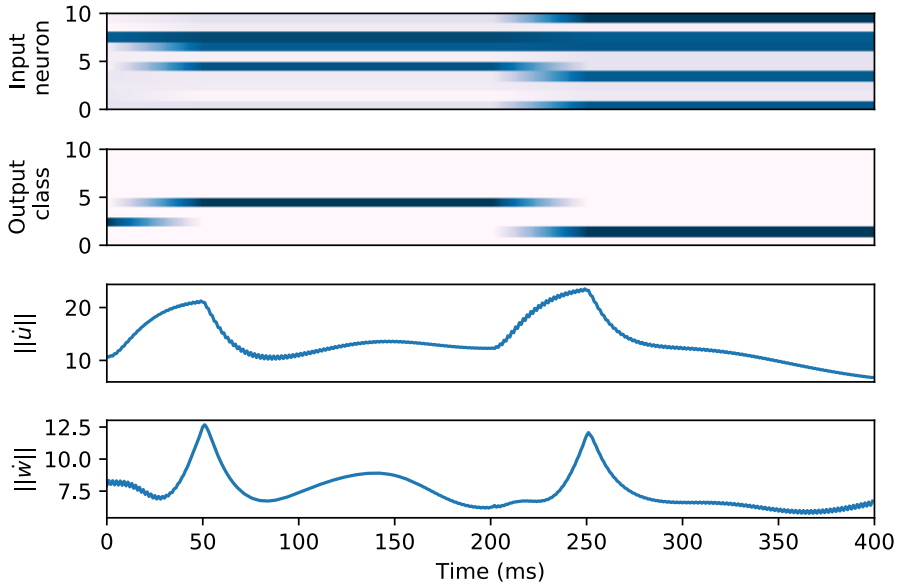


- Make the energy fn of both state s and velocity $\dot{s}$

$$\frac{\partial E(s,\dot{s})}{\partial(s,\dot{s})} = 0$$ can be satisfied all the time

- Hence we can make updates all the time, with the network constantly being nudged towards a better output, no need for phases or waiting for lengthy convergence





Avoiding lengthy convergence: continuous update of the weights



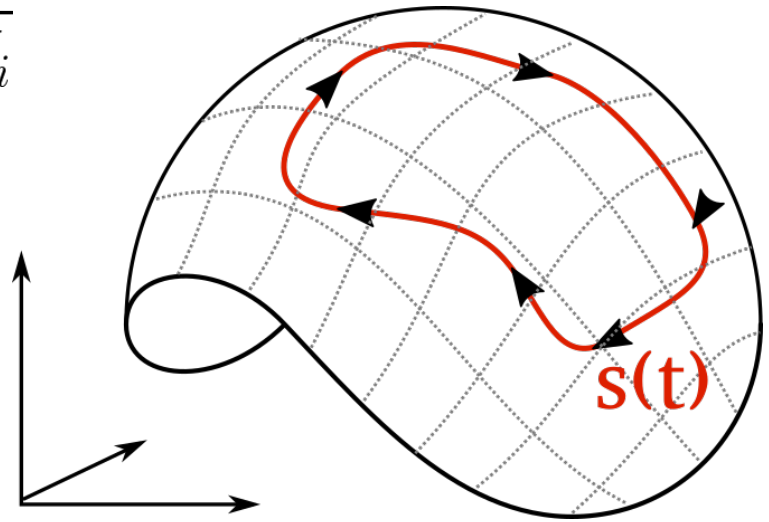
With W. Senn, J. Binas, J. Sacramento

- Rather than deriving dynamics from energy-descent, define dynamics based on Lagrangian mechanics (stationarity of a functional, rather than minimization),

$$\frac{\partial L(q(t),\dot{q}(t),t)}{\partial q_i} - \frac{d}{dt}\frac{\partial L(q(t),\dot{q}(t),t)}{\partial \dot{q}_i} = 0 \text{ for } i = 1,\ldots,n$$

- At the same time, gradient descent on the system energy leads to an update rule compatible with a system composed of interneuron-based microcircuits,

$$\dot{W}_i = -\eta \frac{\partial E}{\partial W_i}$$

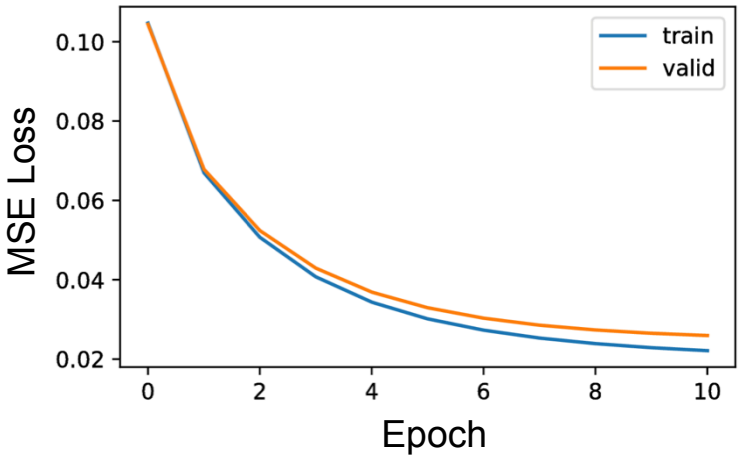


The dynamics is constrained to a manifold, but does not need to converge to a point attractor.

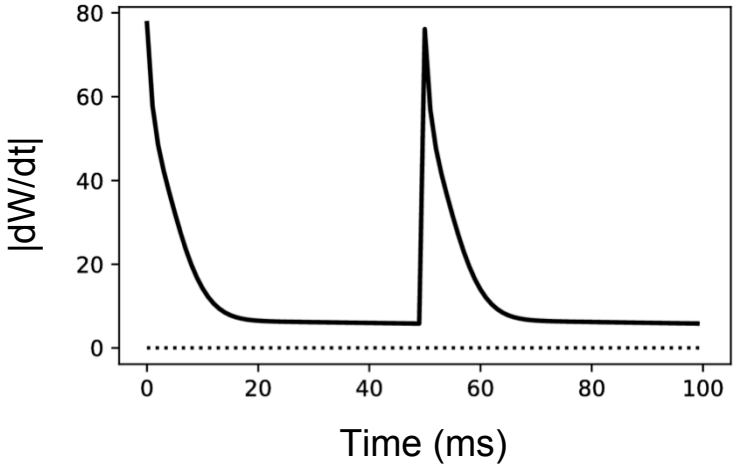
Example model: classifying input patterns in continuous time

MLP: 392-512-512-10

Synthetic random data

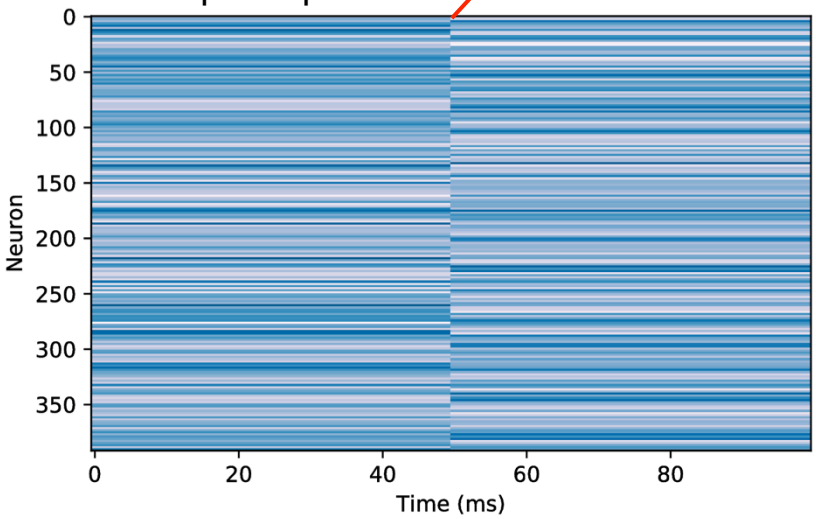


Continuous weight change

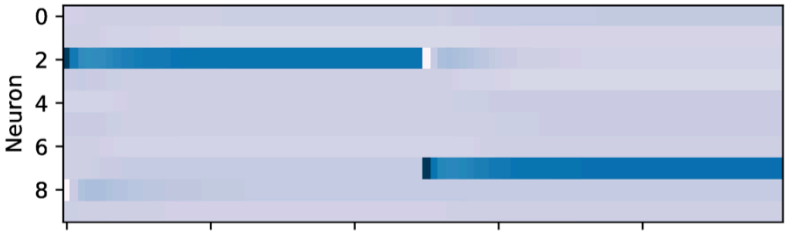


Input pattern switched

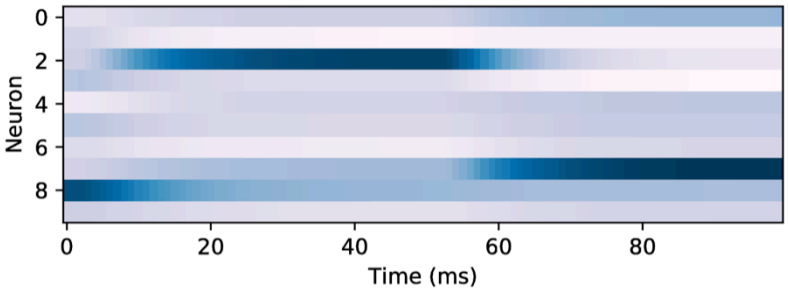
Clamped input



Nudged output



Free-running output



Testing these Theories on real Brains

Need to test on non-toy tasks, which involve 'deep' computation

- Key neuron activity should 'improve' at next trial (gradient descent)
- Key neuron activity should 'improve' 100ms after surprise (back-propagation → nudging)
- Feedback and lateral connections' effects cancel each other on the apical dendrite (mirror interneurons)
  - No surprise → no contribution from apical dendrite
  - Extra stimulation of mirror interneurons → less LTP while reducing the activity of the mirror interneurons → more LTP
- …

Conclusions

- Backprop is the workhorse of the amazing successes of deep learning
- A functionally equivalent implementation of backprop in the brain would help understand its ability to jointly learn in many areas
- No need for a separate net for backpropagated errors, can handle recurrent dynamics and linearity of backprop thanks to mirror interneurons, and no need for 2 separate phases
- Extended to continuous learning with no need to wait for f.p.
- Need extensions to unsupervised learning, models of joint distr.
- Generalize to other energy fns, lateral connections, memory, etc.

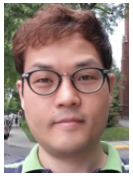
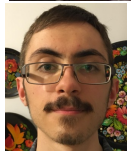
Thanks

Collaborators 2015-2017

- Benjamin Scellier
- Walter Senn
- Joao Sacramento
- Asja Fischer
- Thomas Mesnard
- Dong-Hyun Lee
- Olexa Bilaniuk
- Jonathan Binas
- Thomas Mesnard
- Gaétan Vignoud





Inspiration

Geoff Hinton

MILA