Credit Assignment: Beyond Backpropagation

Yoshua Bengio

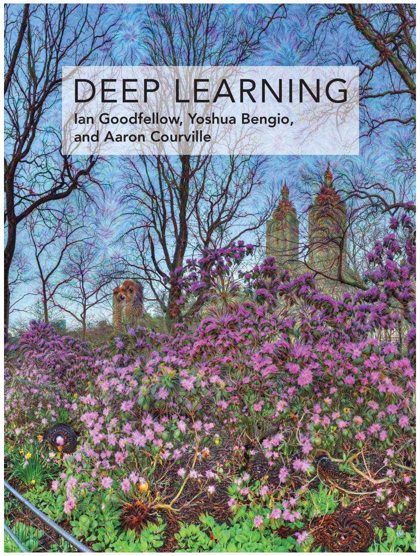
11 December 2016

AutoDiff NIPS'2016 Workshop

MILA

CIFAR CANADIAN INSTITUTE FOR ADVANCED RESEARCH

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



Université de Montréal

IVADO

Deep Learning Jobs in Montreal

- Faculty positions at all levels at U. Montreal
- Researcher positions at Element AI and Google Brain Montreal
- Researcher positions at U. Montreal (IVADO data science center)
- Studentships at all levels at U. Montreal

Something BIG is happening in Montreal

COME AND SEE OUR NEW DIGS AT ivado.ca

IVADO
HEC Montréal
Polytechnique Montréal
Université de Montréal

ELEMENT AI

Some Credit Assignment Principles

- **Chain rule & Backprop**: exact gradient wrt parameters, via gradient wrt intermediate states
  - not always computable, or 0 when discrete operations
  - only valuable in infinitesimal ball
  - but can be stochastic (noise viewed as an extra input)
  - requires storing the full forward computation state (alternative: **forward accumulation**, both memory and compute heavy)
- **Boltzmann machines**: stochastic gradient estimator involves sampling from MCMC, which may have high variance, iterative relaxation
- **REINFORCE** (or random perturbations / finite differences): very general but very high variance, bad scaling
- **Actor-Critic**: trades off some variance for potentially high bias

Backprop wins

- In practice, when backprop can be used, it tends to work MUCH better than any of the other principles
- It can be enhanced by various adaptive techniques (adaptive learning rate, natural gradient, momentum-like techniques)
- Why?
  - only needs to consider ONE direction in the space of variations of the parameters (the gradient)
  - efficient and exact computation of the gradient

Limitations of backprop

- When the computation is DISCRETE or just VERY NONLINEAR we are in trouble with backprop
  - very deep composition of non-linearities, very deep nets (non-ResNet)
  - RNNs with long sequences, problem with long-term dependencies, for the same reason
- **The effect of an infinitesimal change does not always tell us what a small but finite change would yield**

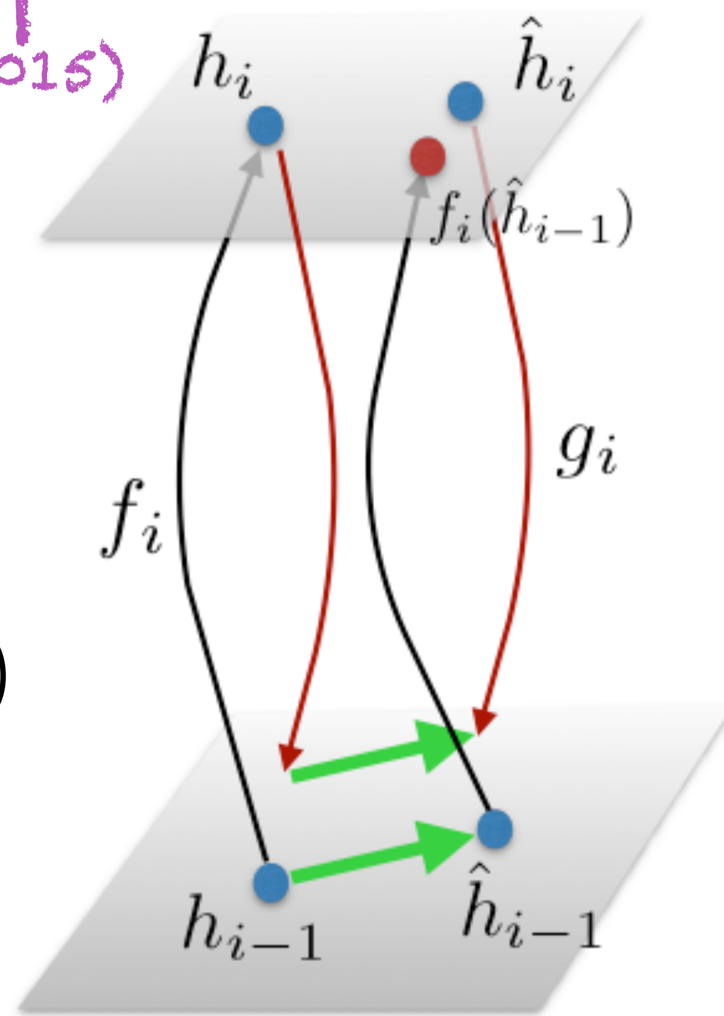
Issues with Boltzmann Machines (with the existing learning procedures)

- Sampling from the MCMC of the model is required in the inner loop of training
- As the model gets sharper, mixing between well-separated modes st

Difference Target-Prop

(Lee, Zhang, Fischer & Bengio 2014 & 2015)

- Make a correction that guarantees to first order that the projection estimated target is closer to the correct target than the original value



$$\hat{h}_{i-1} = h_{i-1} - g_i(h_i) + g_i(\hat{h}_i)$$

$$\left\|\hat{h}_i - f_i(\hat{h}_{i-1})\right\|^2 < \left\|\hat{h}_i - h_i\right\|^2$$

if $1 > \mathit{max\ eigen\ value}\left[(I - f_i'(h_{i-1})g_i'(h_i))^T (I - f_i'(h_{i-1})g_i'(h_i))\right]$

Mostly material from:

Equilibrium Propagation

Bridging the Gap Between Energy-Based Models and Backpropagation

arXiv:1602.0519

Benjamin Scellier & Yoshua Bengio
Montreal Institute for Learning Algorithms

How could we train a continuous time physical system that performs computations?

- Consider a physical system that performs potentially useful computations through its deterministic or stochastic dynamics
- It has parameters θ that could be tuned
- Tractable cost function C can measure how good are its answers
- The relationship between parameters and objective J (cost at equilibrium of the dynamics) is implicit (via the dynamics)
- How to estimate the gradient of the loss wrt parameters?

Equilibria of the Dynamics

- Deterministic case: dynamics converge to fixed points which are minima of an UNKNOWN **energy function** *F*

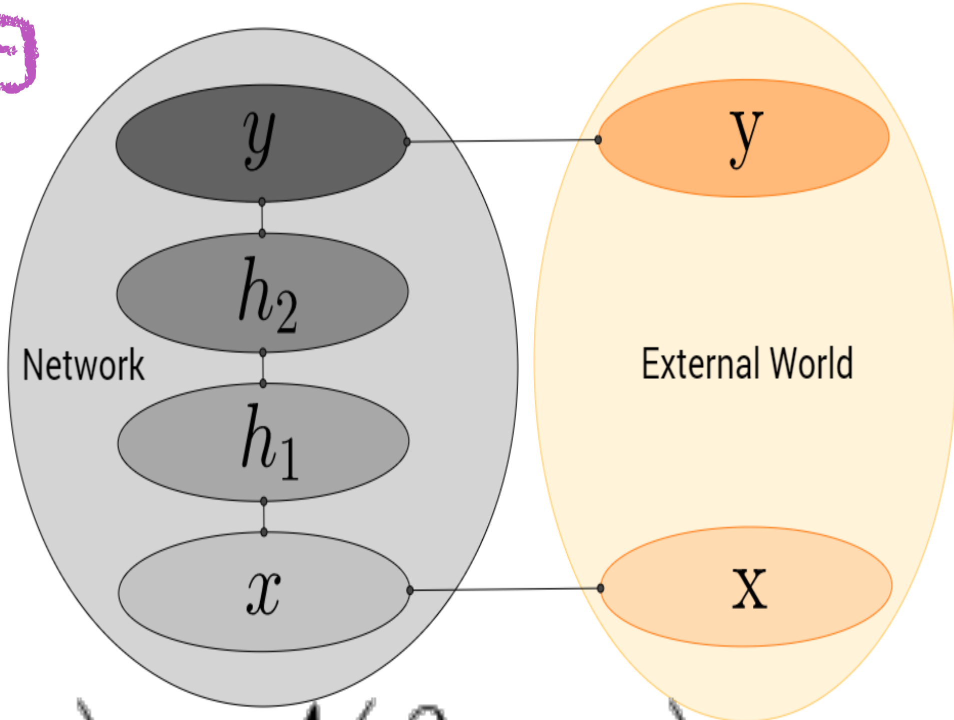
$$\frac{\partial F}{\partial s} = 0 \leftrightarrow \dot{s} = 0$$

- Stochastic case: dynamics converge in probability to the Boltzmann distribution associated with UNKNOWN *F*

$$s \sim P(s) \propto e^{-F(s)}$$

Clamping & Nudging

- The outside world can exert some influence on *v*
- Coefficients β control how much pressure is put on different elements of *v*



$$F(\theta, \beta, s, \mathrm{v}) = E(\theta, s) + A(\beta, s, \mathrm{v})$$

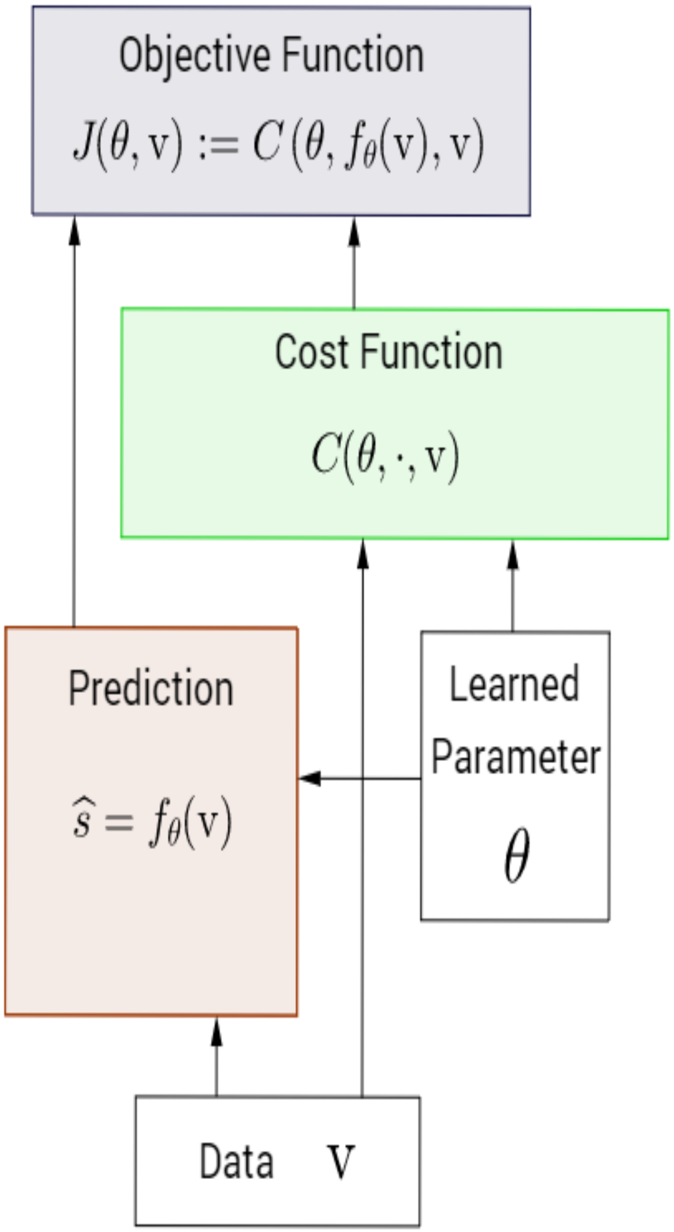
Internal energy

External energy

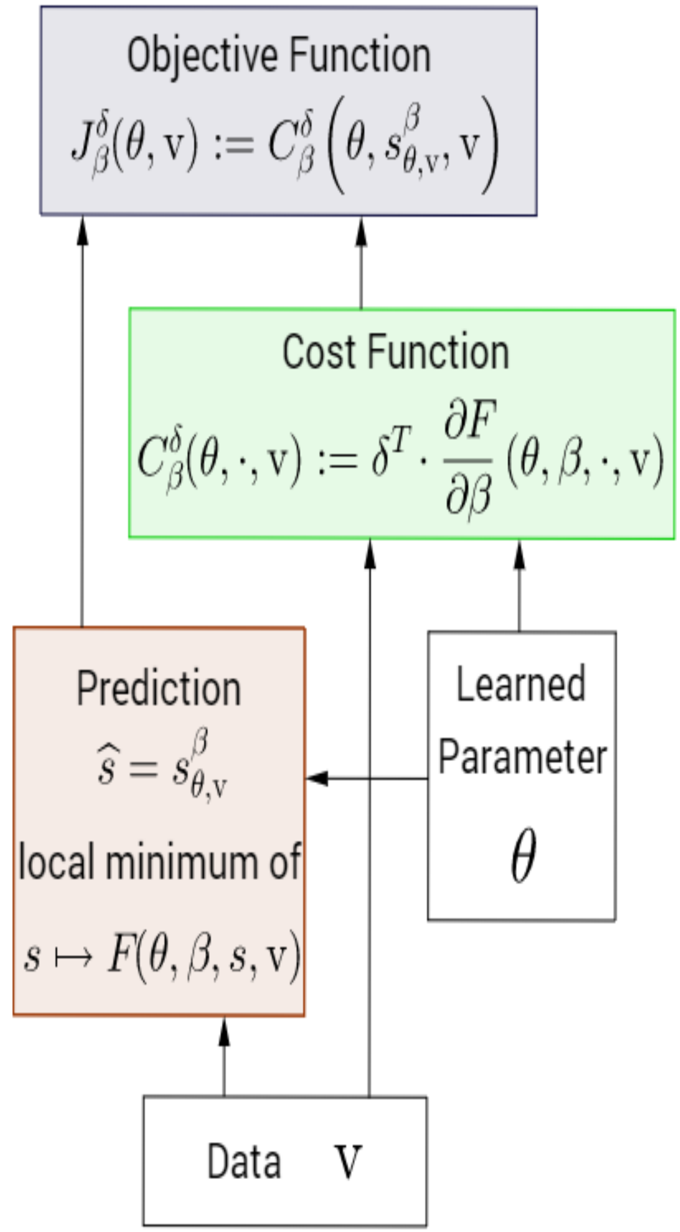
- E.g. $A(\beta, s, \mathrm{v}) = \frac{1}{2}\beta_x||x - \mathrm{x}||^2 + \frac{1}{2}\beta_y||y - \mathrm{y}||^2$

- Clamping *x*: $\beta_x = \infty$ Prediction: clamp *x* and let *y* free with $\beta_y = 0$
- Nudge *y* towards right answer with small $\beta_y = \epsilon$

Traditional ‘Explicit’ Framework



Proposed ‘Implicit’ Framework



Main Theorem

- Gradient on the objective function (cost at equilibrium) can be obtained by a ONE-DIMENSIONAL finite-difference

$$\frac{d}{d\theta} J^{\delta}_{\beta}(\theta, \mathrm{v}) = \lim_{\xi \to 0} \frac{1}{\xi} \left(\frac{\partial F}{\partial \theta} \left(\theta, \beta + \xi\delta, s^{\beta+\xi\delta}_{\theta,\mathrm{v}}, \mathrm{v} \right) - \frac{\partial F}{\partial \theta} \left(\theta, \beta, s^{\beta}_{\theta,\mathrm{v}}, \mathrm{v} \right) \right)$$

Small nudging

Sufficient statistic after nudging

Sufficient statistic before nudging

Stochastic Version

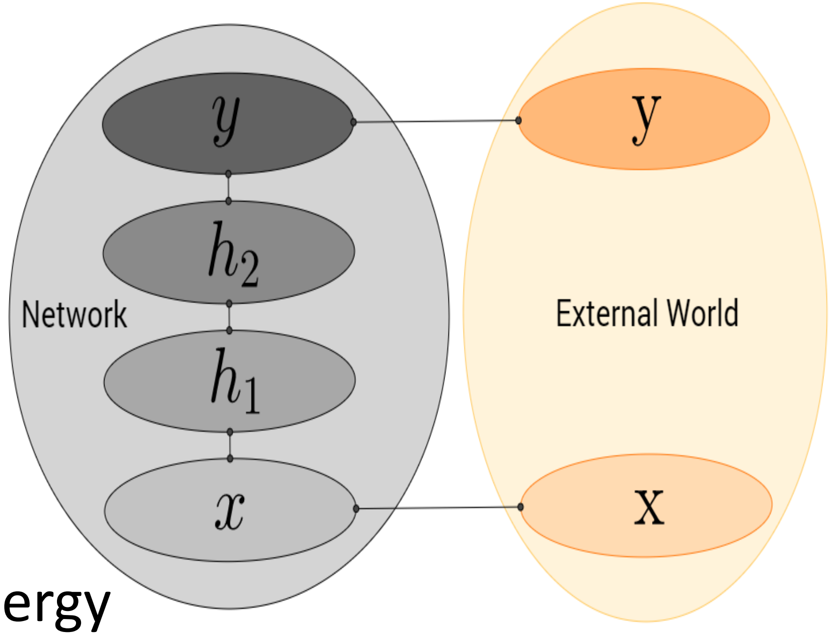
- Equilibrium distribution: $$p^{\beta}_{\theta,\mathrm{v}}(s) := \frac{e^{-F(\theta,\beta,s,\mathrm{v})}}{Z^{\beta}_{\theta,\mathrm{v}}}$$
- Objective = expected cost under that distribution:

$$\widetilde{J}^{\delta}_{\beta}(\theta,\mathrm{v}) := \mathbb{E}^{\beta}_{\theta,\mathrm{v}}\left[\delta \cdot \frac{\partial F}{\partial \beta}(\theta,\beta,s,\mathrm{v})\right]$$

- Theorem:

$$\frac{d}{d\theta}\widetilde{J}^{\delta}_{\beta}(\theta,\mathrm{v}) = \lim_{\xi\to 0}\frac{1}{\xi}\left(\mathbb{E}^{\beta+\xi\delta}_{\theta,\mathrm{v}}\left[\frac{\partial F}{\partial \theta}(\theta,\beta+\xi\delta,s,\mathrm{v})\right] - \mathbb{E}^{\beta}_{\theta,\mathrm{v}}\left[\frac{\partial F}{\partial \theta}(\theta,\beta,s,\mathrm{v})\right]\right)$$

Application to Supervised Learning



- Negative phase:
  - Clamp *x* with $\beta_x = \infty$
  - Let *y* free with $\beta_y = 0$
  - Let dynamics converge to minimum of energy
  - Read out the prediction y and measure loss C
  - Measure sufficient statistics $\dfrac{\partial F}{\partial \theta}$
- Positive phase:
  - Nudge y towards smaller loss by setting $\beta_y = \epsilon$
  - Let dynamics converge to nearby modified min of energy
  - Measure sufficient statistics $\dfrac{\partial F}{\partial \theta}$
- Update parameters towards the difference in suff. stat.

No Need for Calibration of Physical System wrt Idealized Analytic Model

- Traditional analog circuits are meant to approximate an analytic model defined by an equation
- Analog physical implementation are imperfect proxys → need to calibrate and deal with low-precision approximation

- Alternatively: tune the parameters wrt the ACTUAL energy implemented by the physical system, using Equilibrium Propagation

Inherits Properties of Backprop

- Unlike finite-difference methods in parameter space, backprop is equivalent to finite difference IN A SINGLE DIRECTION, THE DIRECTION OF THE COST GRADIENT. Same here.

- In the case where the network has a multi-layer structure, we can show that the propagation of perturbations (nudges) corresponds to back-propagation of gradients
    - First shot at showing this in
        - *Bengio & Fischer, Early Inference in Energy-Based Models Approximates Back-Propagation, arXiv:1510.02777*

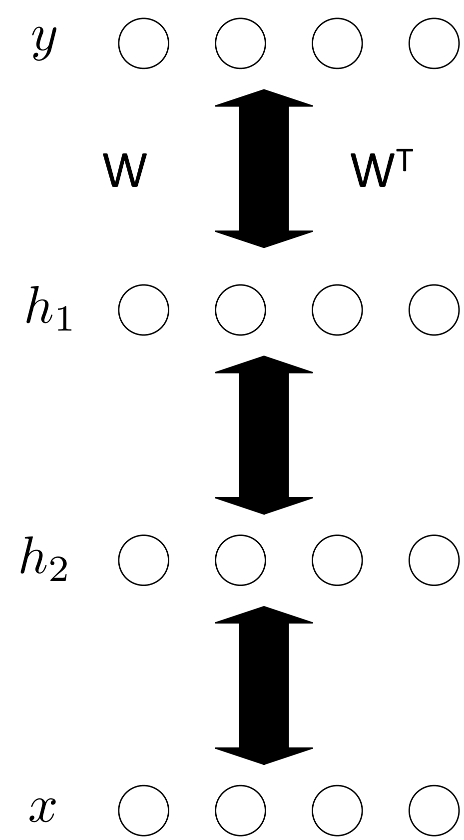
Propagation of errors = propagation of surprises = getting back in harmony

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 $-\frac{\partial C}{\partial h}$



Equilibrium Propagation Includes Ordinary Backprop for Feedforward Nets as Special Case

- Consider the internal energy function

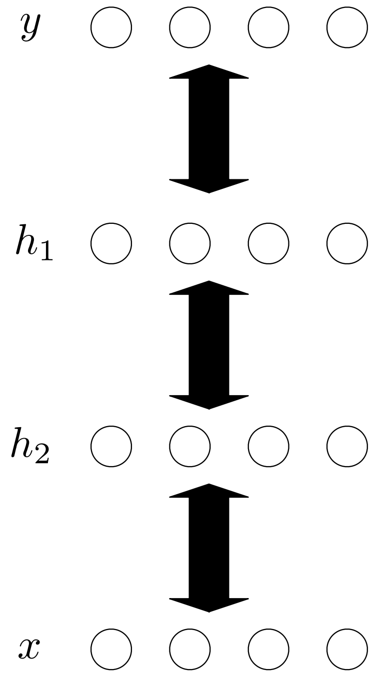
$$E = \sum_l ||h_l - f_l(h_{l-1})||^2$$

With layered architecture, h_l= *l*-th layer of activations, $h_0 = x$

f_l = parametrized computation at *l*-th layer.

- E has a global minimum at $h_l = f_l(h_{l-1})$
- It is also a mode associated with stationary distribution.

Equilibrium Propagation Includes Ordinary Backprop for Feedforward Nets as Special Case



- With this feedforward-compatible energy-function

$$E = \sum_l ||h_l - f_l(h_{l-1})||^2$$

- Negative phase is EQUIVALENT to feedforward prop.
- Positive phase: nudge outputs, nudges propagated backwards
- Equilibrium-propagation estimates the same gradient as backprop in a feedforward net, but using a physical (analog) dynamical system which implements the above energy function, with no need for a separate circuit for backpropagation.

Connection to Marginal Log-Likelihood (Boltzmann Machine)

- Define $\phi(\xi) = \log E_{p_\beta}[e^{-\xi C}]$
- Thm: if *F* is linear in β then

$$\phi(\xi) = \log Z_{\beta+\xi\delta} - \log Z_\beta$$

- Note that $\frac{\partial \log Z_\beta}{\partial \theta} = -E_{p_\beta}[\frac{\partial F}{\partial \theta}]$ $\quad \phi'(0) = J_\beta$
- Corollary:

$$\frac{\partial \phi(\xi)}{\partial \theta} = -E_{p_{\beta+\xi\delta}}[\frac{\partial F(\theta, \beta+\xi\delta, s, v)}{\partial \theta}] + E_{p_\beta}[\frac{\partial F(\theta, \beta, s, v)}{\partial \theta}]$$

$$\lim_{\xi \to \infty} \frac{\partial \phi(\xi)}{\partial \theta} = -\frac{\partial \log p_\beta(v)}{\partial \theta}$$

Interpretation for Biological Implementation of Backprop → STDP

- This was the initial motivation
- Hopfield(-like) energy function

$$E(s) = \sum_i \frac{s_i^2}{2} - \frac{1}{2} \sum_{i \neq j} W_{i,j} \rho(s_i) \rho(s_j) - \sum_i b_i \rho(s_i)$$

- gives rise to neurally plausible dynamics (with gradient descent or Langevin dynamics) $-\frac{\partial E}{\partial s_i}(\theta, s) = \rho'(s_i) \left(\sum_{j \neq i} W_{ij} \rho(s_j) + b_i \right) - s_i$
- Sufficient statistics = Hebbian $\frac{\partial E}{\partial W_{ij}}(\theta, s) = -\rho(s_i)\rho(s_j)$
- Update: a form of contrastive Hebbian update

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

- Can be implemented by continuously following $\frac{d}{dt}\rho(s_i)\rho(s_j)$ during the positive phase = STDP update.
- Remaining issue: need symmetric weights.

Some Open Problems

- How to implement this in analog electric circuit? With a voltage source and current flowing, there is no equilibrium in terms of electrons' energy (position & momentum), only in terms of currents and voltages: Lyapunov function?
- Get rid of local minima of energy formulation and generalize to system defined purely by its dynamics, learn the transition operator, thus avoiding the weight symmetry constraint
- Generalize these ideas to unsupervised learning (ongoing)


Montreal Institute for
Learning Algorithms
DARTH
VADER


MILA
Université
de Montréal