

Towards Biologically Plausible Deep Learning

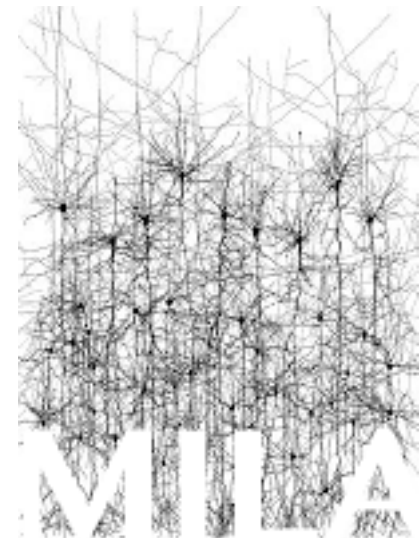
Yoshua Bengio

March 12, 2015

NYU

*Yoshua Bengio, Dong-Hyun Lee, Jorg Bornschein, and Zhouhan
Lin, arXiv 1502.04156*

Université 
de Montréal

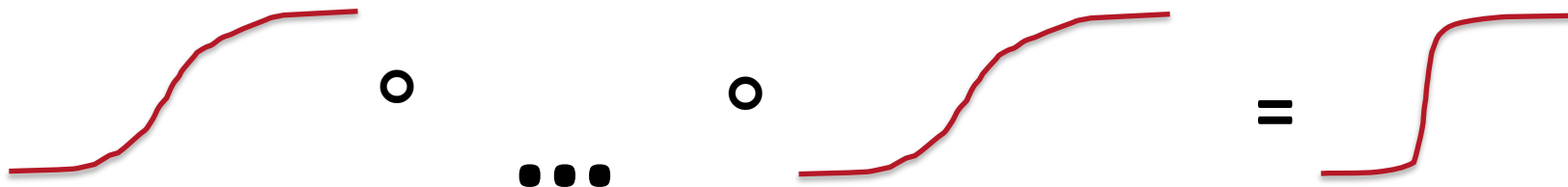


Central Issue in Deep Learning: Credit Assignment

- What should hidden layers do?
- Established approaches:
 - Backpropagation
 - Stochastic relaxation in Boltzmann machines

Issues with Back-Prop

- Over very deep nets or recurrent nets with many steps, non-linearities compose and yield sharp non-linearity \rightarrow gradients vanish or explode
- Training deeper nets: harder optimization
- In the extreme of non-linearity: discrete functions, can't use back-prop
- Biological plausibility



Biological Plausibility Issues with Standard Backprop

1. BP of gradient = purely linear computation, not plausible across many neural levels
2. If feedback paths are used for BP, how would they know the precise derivatives of forward-prop?
3. Feedback paths would have to use exactly the same weights (transposed) as feedforward paths
4. Real neurons communicate via spikes
5. Need to clock and alternate feedforward and feedback computation
6. Where would the supervised targets come from?

Issues with Boltzmann Machines

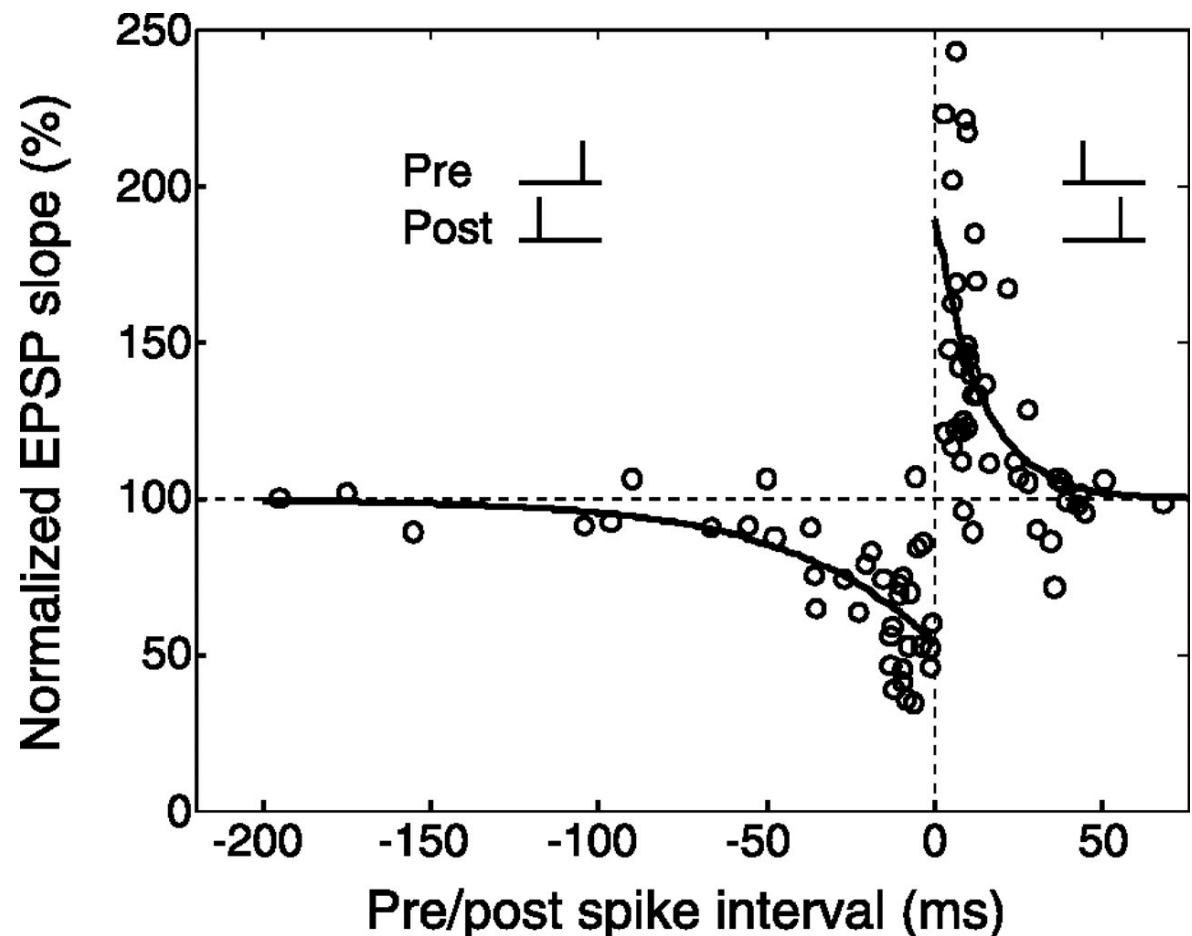
- Sampling from the MCMC of the model is required in the inner loop of training
- As the model gets sharper, mixing between well-separated modes stalls



What is the brain's Learning algorithm?

Cue: Spike-Timing Dependent Plasticity

- Observed throughout the nervous system, especially in cortex
- STDP: weight increases if post-spike just after pre-spike, decreases if just before.



Machine Learning Interpretation of Spike-Timing Dependent Plasticity

- Suggested by Xie & Seung NIPS'99 and Hinton 2007: the STDP update corresponds to a temporal derivative filter applied to post-spike, around pre-spike.
- In agreement with the above, we argue this corresponds to

$$\Delta W_{ij} \propto S_i \Delta V_j$$

synaptic change pre-spike temporal change in post-potential

Machine Learning Interpretation of Spike-Timing Dependent Plasticity

$$\Delta W_{ij} \propto S_i \Delta V_j$$

- would be SGD on objective J if

$$\Delta V_j \approx \frac{\partial J}{\partial V_j}$$

- This corresponds to neural dynamics implementing a form of inference wrt J , seen as a function of parameters and latent vars

STDP and Variational EM

- Neural dynamics moving towards “improved” objective J and parameter updates towards the same J corresponds to a variational EM learning algorithm,

$$\log p(x) \geq E_{q^*(H|x)} [\log p(x, H)]$$

Approximate inference

- where J = regularized joint likelihood of observed x and latent h

$$J = \log p(x, h) + \text{regularizer}$$

Generative model /
All interactions between neurons

Inference initial guess
(forward pass)

- Generalizes PSD (Predictive Sparse Decomposition) from (Kavukcuoglu & LeCun 2008) with regularizer $= \alpha q(h|x)$

What Inference Mechanism?

- Simply going down on J 's gradient corresponds to MAP inference (disadvantage: decoder not sufficiently contractive)
- Injecting noise in the process gives a form of approximate posterior MCMC, such as Langevin MCMC

$$\dot{h} = \frac{1}{2\sigma} \frac{\partial J}{\partial h} + \sigma \text{ Brownian noise}$$

- Or, in discrete time:
$$h \leftarrow h + \frac{1}{2\sigma} \frac{\partial J}{\partial h} + \sigma \text{ Normal}(0, 1) \text{ noise}$$

* no rejection: biased samples, but ok, *see (Welling & Teh ICML 2011)*

Inference Decouples Deep Net Layers

- After inference, no need for back-prop because the joint over layers decouples the updates of the parameters from the different layers, e.g.

Generative model \rightarrow $p(x, h) = p(x|h^{(1)}) \left(\prod_{k=1}^{M-1} p(h^{(k)}|h^{(k+1)}) \right) p(h^{(M)})$

Parametric initialization for approximate inference \rightarrow $q(h|x) = q(h^{(1)}|x) \prod_{k=1}^{M-1} q(h^{(k+1)}|h^{(k)})$

- So J could be of the form

$$J = \sum_k \log p(h^{(k)}|h^{(k+1)}) + \log q(h^{(k+1)}|h^{(k)})$$

But Inference Seems to Need Backprop

Iterative inference, e.g. MAP

Initialize $h \sim q(h|x)$

for $t = 1$ to T **do**

$$h \leftarrow h + \delta \frac{\partial J}{\partial h}$$

Involves $\frac{\partial J}{\partial h}$ which has terms of the form

$$\frac{\partial \log p(h^{(k-1)} | h^{(k)})}{\partial h^{(k)}}$$

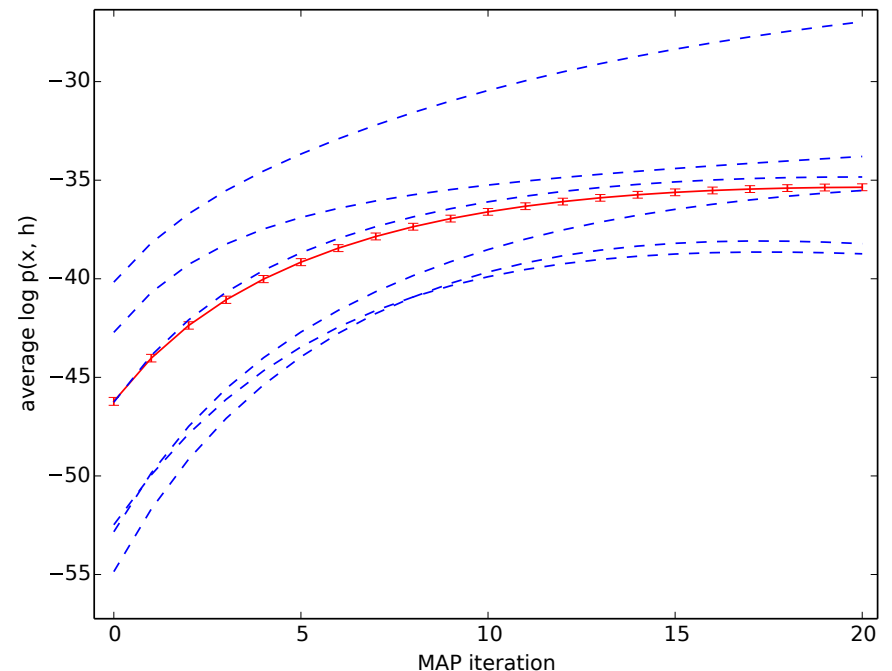
to change upper layer to make lower layer value more probable (or the equivalent for q)

But Inference Seems to Need Backprop

How to back-prop through one layer
without explicit derivatives?

DIFFERENCE *TARGET-PROP*

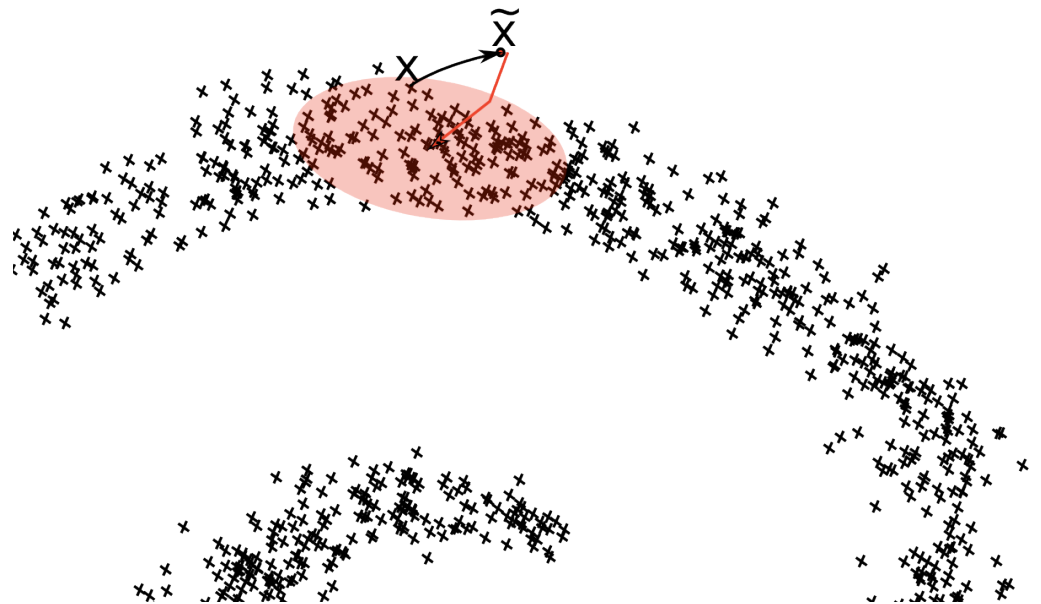
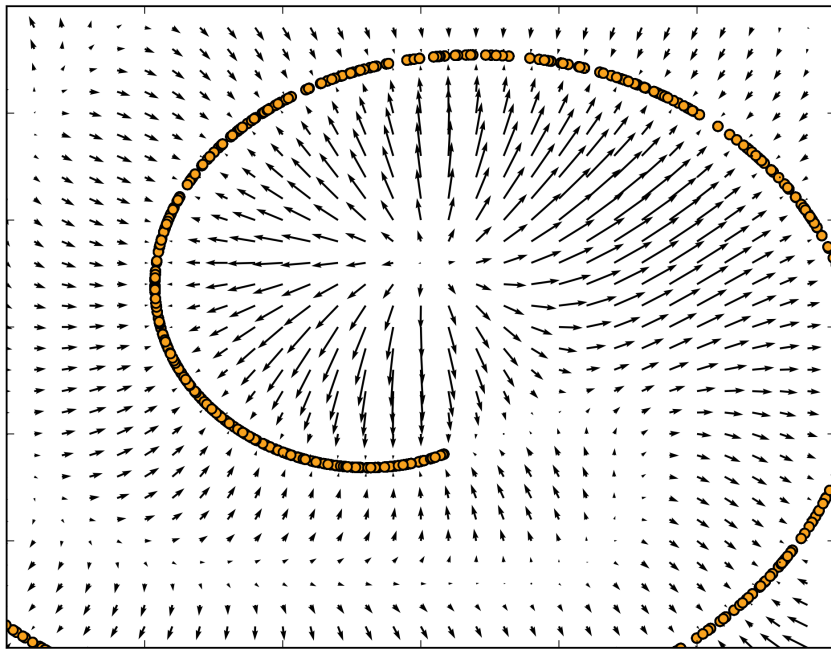
***Result: iterative inference
climbs J even though no
gradients were ever computed
and no animal was harmed!***



Parenthesis about auto-
encoders probabilistic
interpretation

Regularized Auto-Encoders Learn a Vector Field or a Markov Chain Transition Distribution

- (Bengio, Vincent & Courville, TPAMI 2013) review paper
- (Alain & Bengio ICLR 2013; Bengio et al, NIPS 2013)



Denoising Auto-Encoders Learn a Small Move Towards Higher Probability

(Alain & Bengio ICLR 2013)

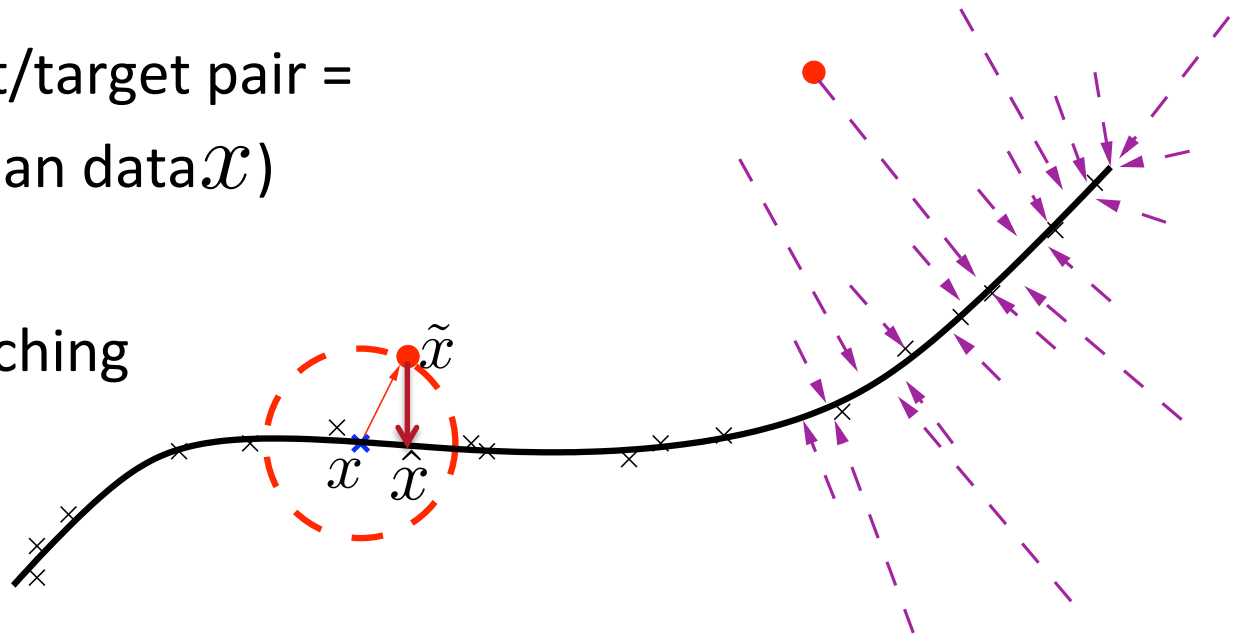
- Reconstruction \hat{x} points in direction of higher probability

$$\hat{x} - x \propto \frac{\partial \log P(x)}{\partial x}$$

gradient

- Trained with input/target pair =
(corrupted $\tilde{x} \rightarrow$ clean data x)

- DAE \rightarrow Score matching
(Vincent 2011)



General Result about Denoising

(Alain & Bengio ICLR 2013)

- Non-parametric limit:

$$r^* = \operatorname{argmin}_r E[||x - r(x + \sigma z)||^2]$$

- where z is $N(0,1)$ noise and $E[.]$ is over $p(x)$ and z . Then


$$\frac{r^*(x) - x}{\sigma^2} = \frac{\partial \log p(x)}{\partial x}$$

- i.e., following the reconstruction goes down the gradient

Consistency Results (Bengio et al NIPS 2013)

- Denoising AE are consistent estimators of the data-generating distribution through their Markov chain (corrupt, reconstruct and inject reconstruction error noise, repeat), so long as they consistently estimate the conditional denoising distribution and the Markov chain converges.

Making $P_{\theta_n}(X|\tilde{X})$ match $\mathcal{P}(X|\tilde{X})$ makes $\pi_n(X)$ match $\mathcal{P}(X)$


denoising distr. truth stationary distr. truth

- In other words, if the inference mechanism corresponds to corruption and denoising reconstruction, we are following the model's Markov chain.

Denoising Score Matching

- An alternative to maximum likelihood for continuous random variables
- Asymptotically consistent estimator (as noises level decreases and # examples increases)
- Reconstruction: $r(x) = x - \sigma^2 \frac{\partial \text{Energy}(x)}{\partial x}$
- Denoising training objective, with $N(0,1)$ noise z :

$$E_{x,z} [\|r(x + \sigma z) - x\|^2]$$

→ No partition function gradient!

Extracting Structure By Gradual Disentangling and Manifold Unfolding

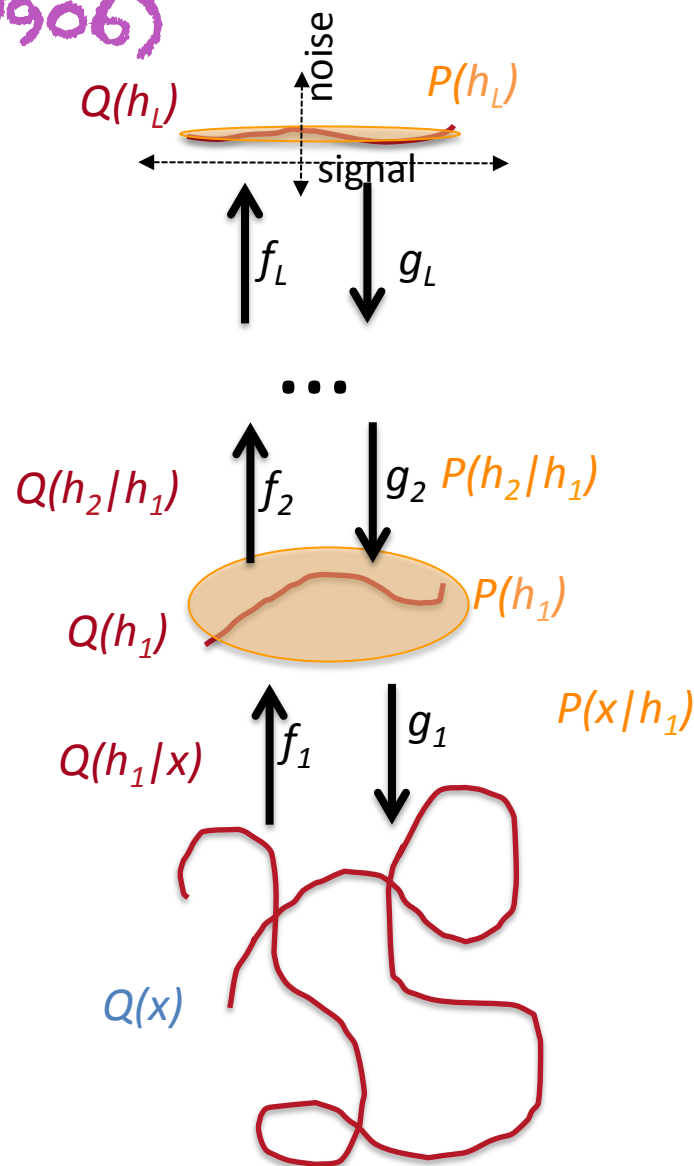
(Bengio 2014, arXiv 1407.7906)

Each level transforms the data into a representation in which it is easier to model, unfolding it more, contracting the noise dimensions and mapping the signal dimensions to a factorized (uniform-like) distribution.

$$\min KL(Q(x, h) || P(x, h))$$

= variational auto-encoder criterion

(Kingma & Welling ICLR 2014)



Close parenthesis

Difference Target-Prop Estimator

- If the encoder is $f(x)$ +noise and the decoder is $g(h)$ +noise, then

$$\frac{\partial \log p(x|h)}{\partial h} \approx \frac{f(x) - f(g(h))}{\sigma_h^2}$$

- which is demonstrated by exploiting

$$\log p(x|h) = \log p(x, h) - \log p(h)$$

- and the DAE score estimator theorem

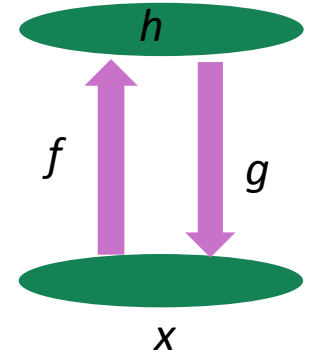
$$\frac{r(x) - x}{\sigma^2} \rightarrow \frac{\partial \log p(x)}{\partial x}$$

- Considering two DAEs, one with h as “visible” and one with (x, h)

Decomposition of the gradient into reconstructions

- We want

$$\frac{\partial \log p(x|h)}{\partial h} = \frac{\partial \log p(x, h)}{\partial h} - \frac{\partial \log p(h)}{\partial h}$$



- which we get from two auto-encoders:

1. The (x, h) to (h, x) AE: $r(x, h) = (g(h), f(x))$

$$\rightarrow \frac{f(x) - h}{\sigma^2} \approx \frac{\partial \log p(x, h)}{\partial h}$$

2. The AE with h as « visible » and x as « representation »

$$\rightarrow \frac{f(g(h)) - h}{\sigma^2} \approx \frac{\partial \log p(h)}{\partial h}$$

- Result:

$$\frac{\partial \log p(x|h)}{\partial h} \approx \frac{f(x) - f(g(h))}{\sigma_h^2}$$

Same Formula justifies Backprop-free Auto-Encoder based on Target-Prop

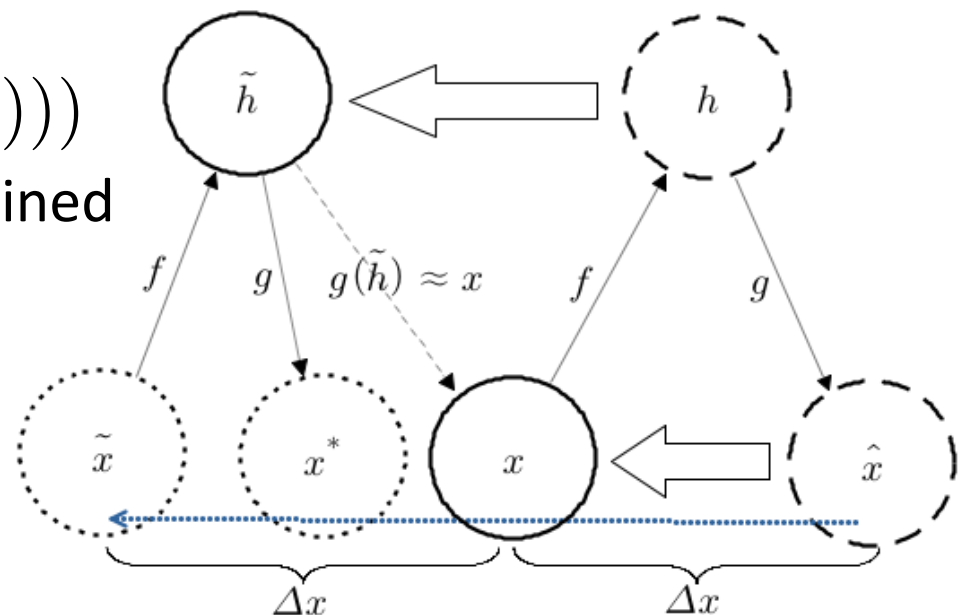
- If $r(x)=f(g(h))$ is smooth and makes a small move away from x , then applying r from

$$\tilde{x} = x - \Delta x = x - (g(f(x)) - x) = 2x - g(f(x))$$

- should approximately give x , so $g(\tilde{h}) \approx x$
- where

$$\tilde{h} = f(\tilde{x}) = f(2x - g(f(x)))$$

- And the encoder should be trained on the pair (\tilde{x}, \tilde{h})



Difference Target-Prop for Inexact Inverse

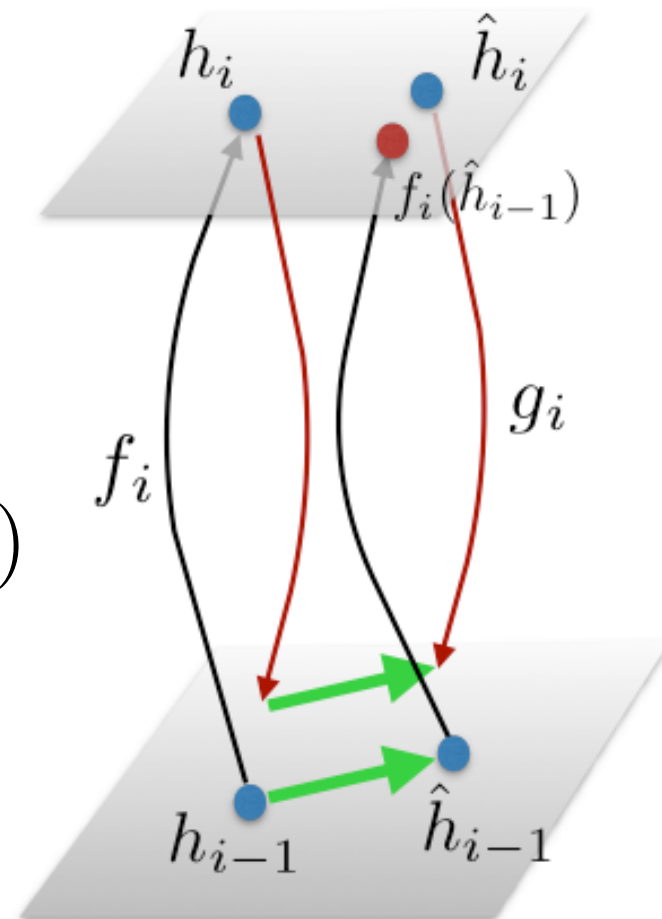
- Make a correction that guarantees to first order that the projection estimated target is closer to the correct target than the original value

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

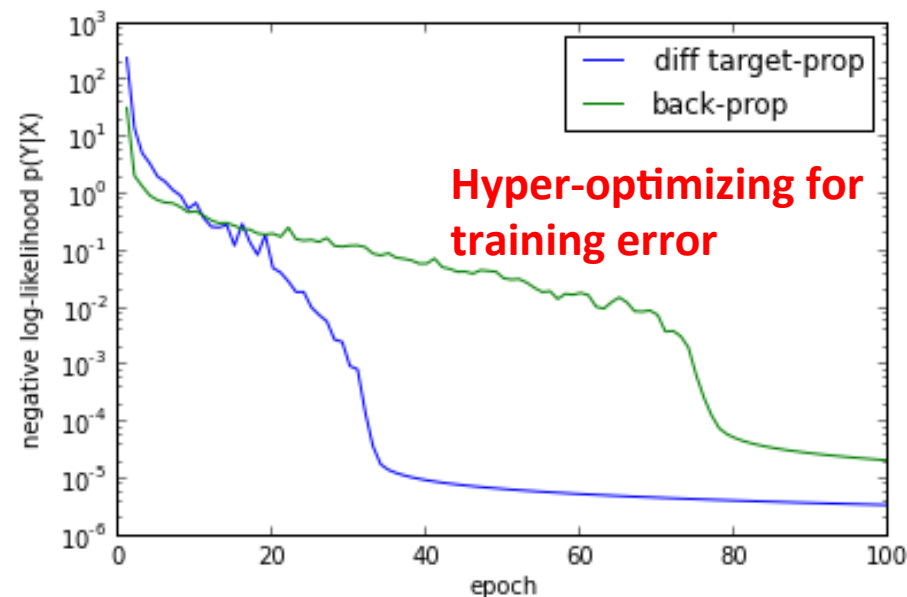
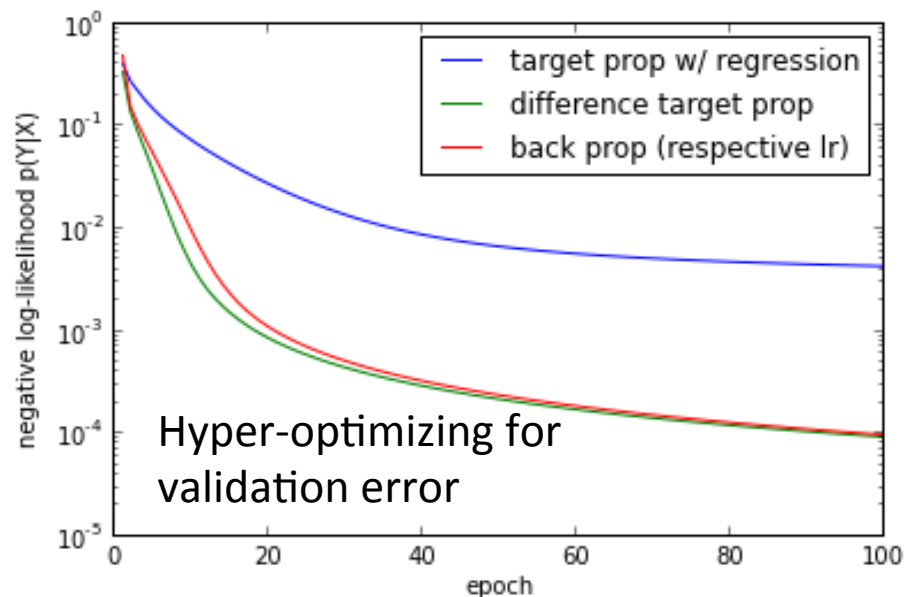
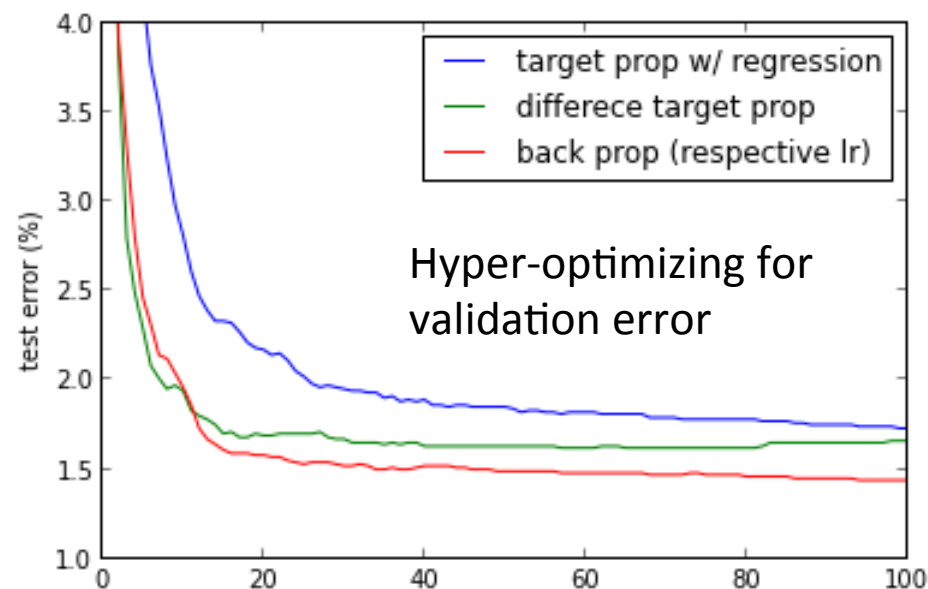
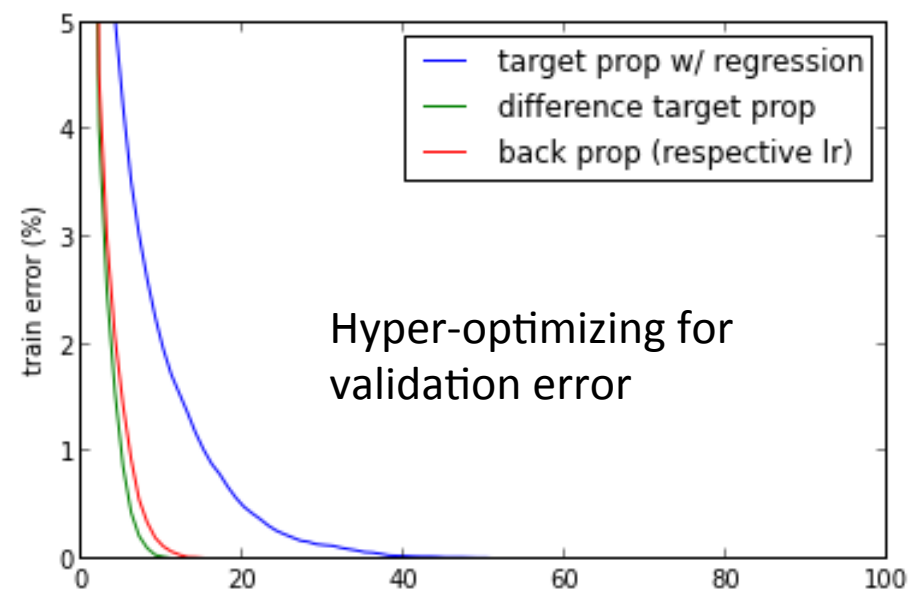
- Special case: feedback alignment, if $g_i(h) = B h$

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

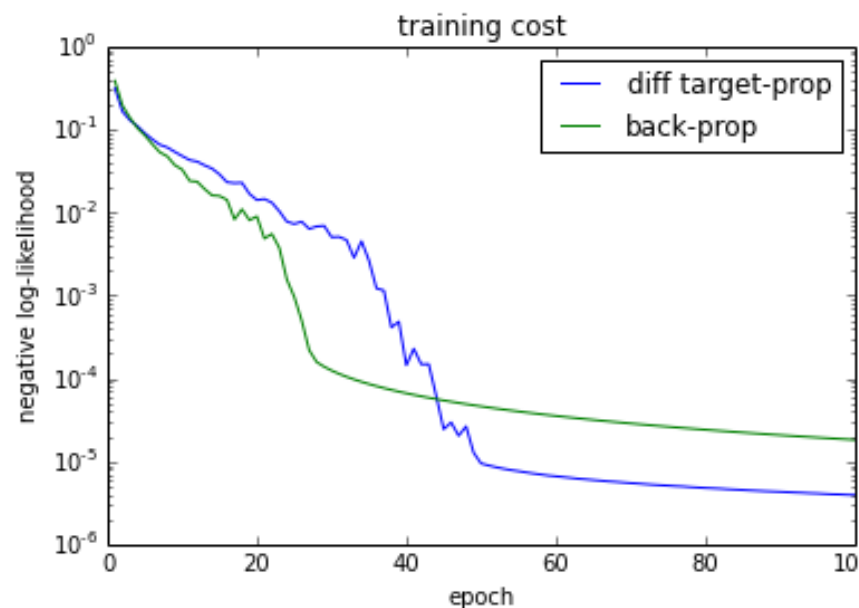
if $1 > \max \text{ eigen value } \left[(I - f'_i(h_{i-1})g'_i(h_i))^T (I - f'_i(h_{i-1})g'_i(h_i)) \right]$



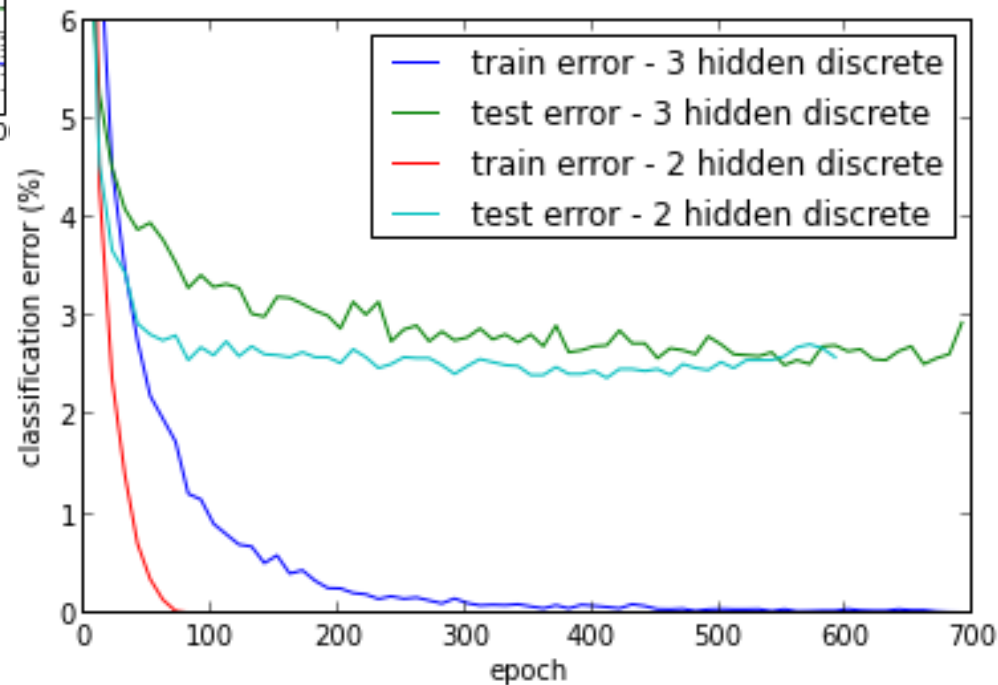
Obligatory MNIST Results (supervised target-prop)



Targetprop can work for discrete and/or stochastic activations



Work in progress

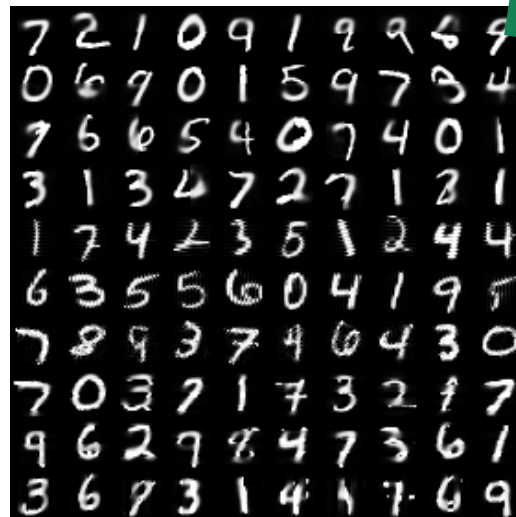


Iterated Target-Prop Generative Deep Learning Experiments on MNIST

Generated model samples



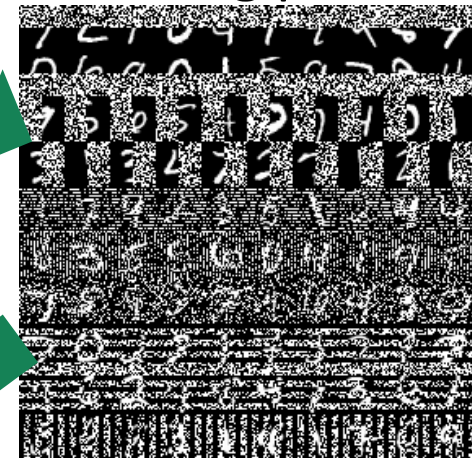
Inpainted



Original examples



Inpainting starting point



Inpainting missing values (starting from noise)

What's Next?

- Experiments only involved p terms in J , but if there is going to be multiple modalities, we need correction signals (target prop) from above as well as from below
- Using true gradients instead of diff targetprop yielded better final values of J after each inference iteration but a worse final value of J after training. Why?
- Proposed theory suggests that using only a few inference iterations should give a sufficient signal to update weights, but experiments required 10-15.
- Updates in paper did not follow the STDP framework but used final inference values as targets

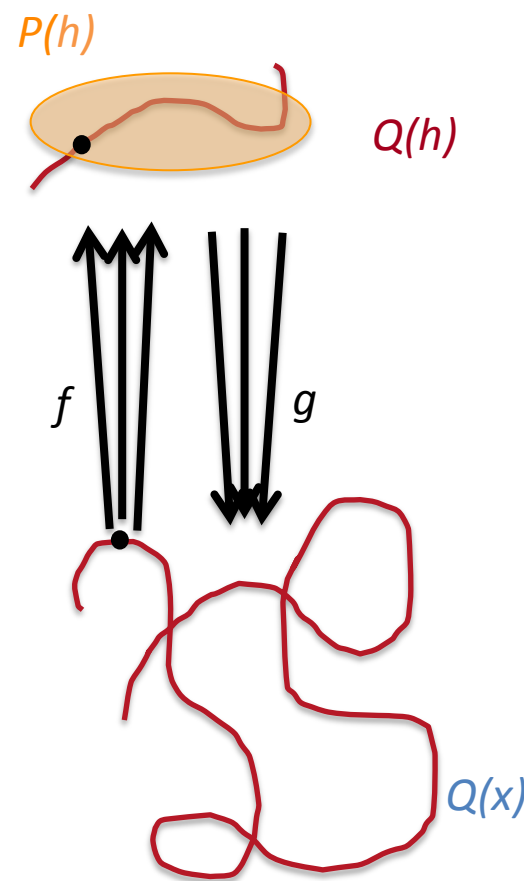
Why Noise is Needed

- Up to now we used a MAP inference in our experiments
- Adding noise appropriately makes it a biased Langevin MCMC, making the inference procedure approximately sample from the posterior of latent given visible
- Noise may be **necessary** to appropriately prepare the decoder to face the inadequacy of the higher-levels 'prior', by becoming contractive
- It comes up automatically in the variational auto-encoder criterion

The Importance of Contractive Decoder

- Denoising \rightarrow contractive g
- Max. determinant of $f' \rightarrow f$ expansive at data x , g contractive around
- Contraction \rightarrow removes unnecessary directions
- Making g contractive helps to manage the mismatch between $P(h)$ and $Q(h)$
- Adding noise at the top-level in $Q(h/x)$ shows to the decoder which directions of h need to be contracted out, making it contractive

If f bijective $P(x) = P(h=f(x)) / |\det f'(x)|$



Many Probabilistic Interpretations e.g. EM Denoising Score Matching

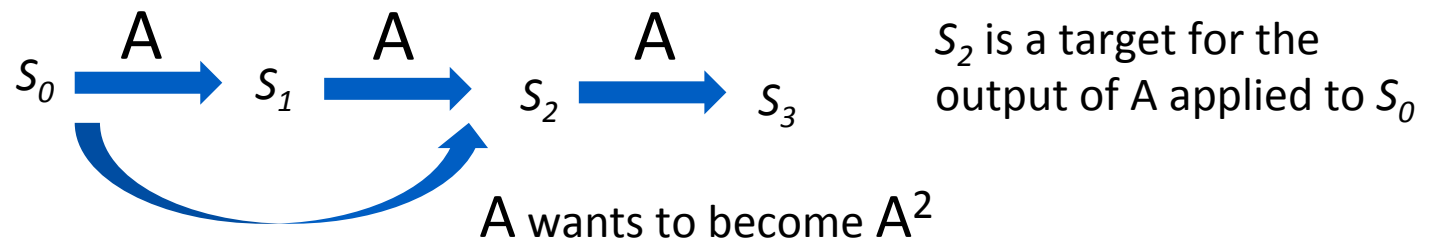
- A reconstruction function (state \rightarrow state) embodies energy gradient (to improved state) and defines neural dynamics
- Use it for inference, e.g. Langevin MCMC, i.e., update state towards reconstruction, with some noise injected
- Given visible x , do inference to sample $h \sim$ posterior given x
- Consider state $s=(x,h)$ as if they were visible and perform a denoising score matching update of parameter i.e.,

$$\min_{\theta} ||\text{reconstruct}(\text{corrupt}(\text{state})) - \text{state}||^2$$

- Any energy function can be defined, but some give rise to biologically plausible neural dynamics

Ongoing: Impatient Learned Approximate Inference

- Instead of waiting for the last step of inference (to be used as target a la EM), we can ask each inference step to land where the next step will land, i.e., to speed-up the MCMC burn-in
- i.e., target state = later in the chain
corrupted state = noisy, earlier state in the chain
reconstruction error becomes PREDICTION error
- This would result in an SDTP-like update, at every time step, not just at the end of inference



MILA: Montreal Institute for Learning Algorithms

