Learning Deep Feature Hierarchies

Yoshua Bengio, U. Montreal

Stanford University, California

September 21st, 2009

Thanks to: Aaron Courville, Pascal Vincent, Dumitru Erhan, Olivier Delalleau, Olivier Breuleux, Yann LeCun, Guillaume Desjardins, Pascal Lamblin, James Bergstra, Nicolas Le Roux, Max Welling, Myriam Côté, Jérôme Louradour, Ronan Collobert, Jason Weston

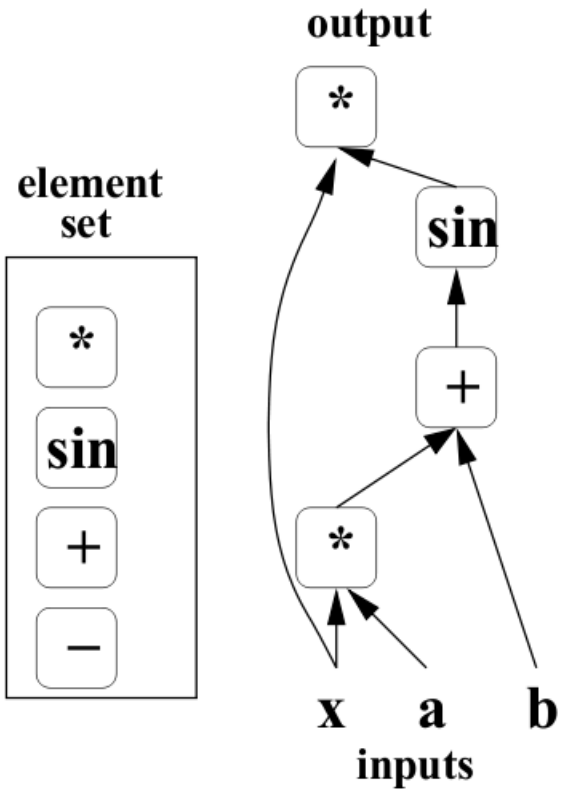
Interesting Experimental Results with Deep Architectures

- Beating shallow neural networks on vision and NLP tasks
- Beating SVMs on visions tasks from pixels (and handling dataset sizes that SVMs cannot handle in NLP)
- Reaching or beating state-of-the-art performance in NLP
- Beating deep neural nets without unsupervised component
- Learn visual features similar to V1 and V2 neurons

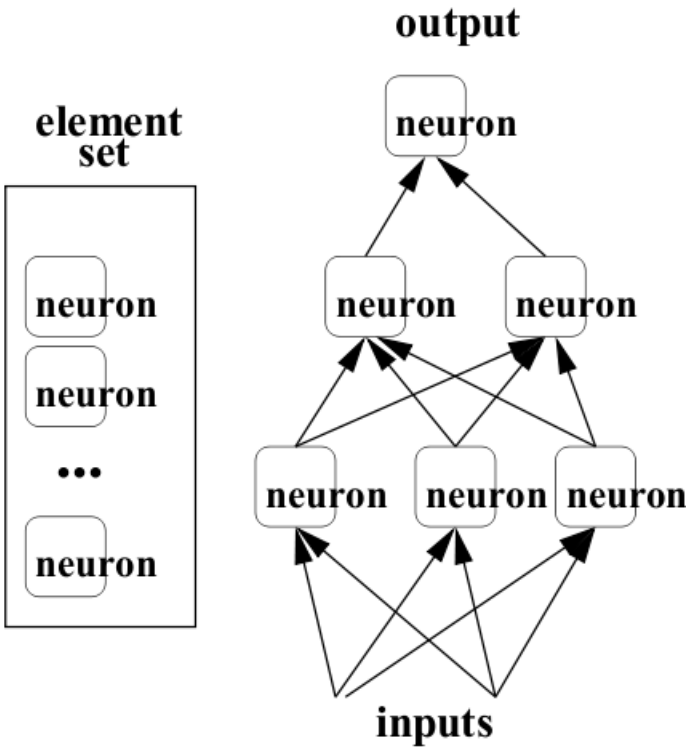
Deep Motivations

- Brains have a deep architecture
- Humans organize their ideas hierarchically, through composition of simpler ideas
- Unsufficiently deep architectures can be exponentially inefficient
- Distributed (possibly sparse) representations are necessary to achieve non-local generalization
- Multiple levels of latent variables allow combinatorial sharing of statistical strength

Architecture Depth



Depth = 4



Depth = 3

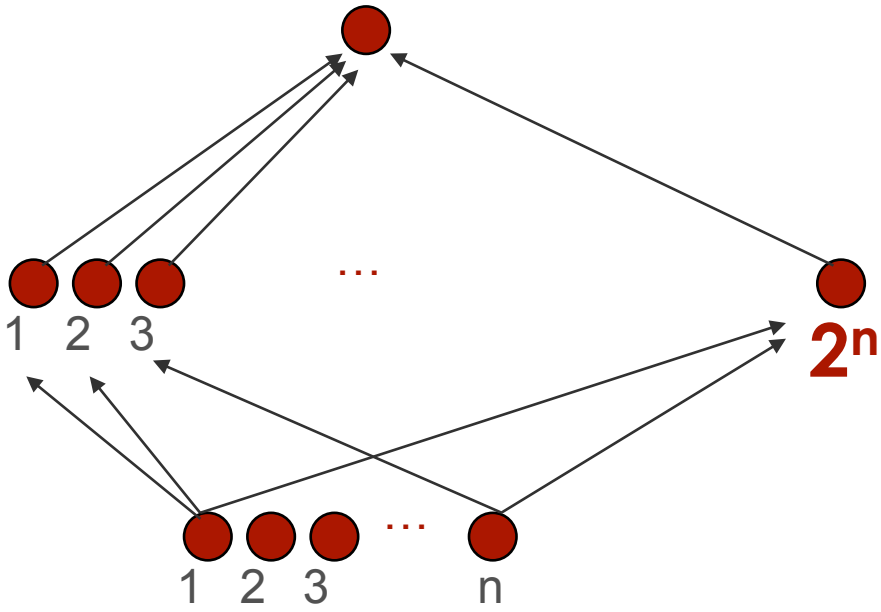
Deep Architectures are More Expressive

Theoretical arguments:

2 layers of { Logic gates / Formal neurons / RBF units } = universal approximator

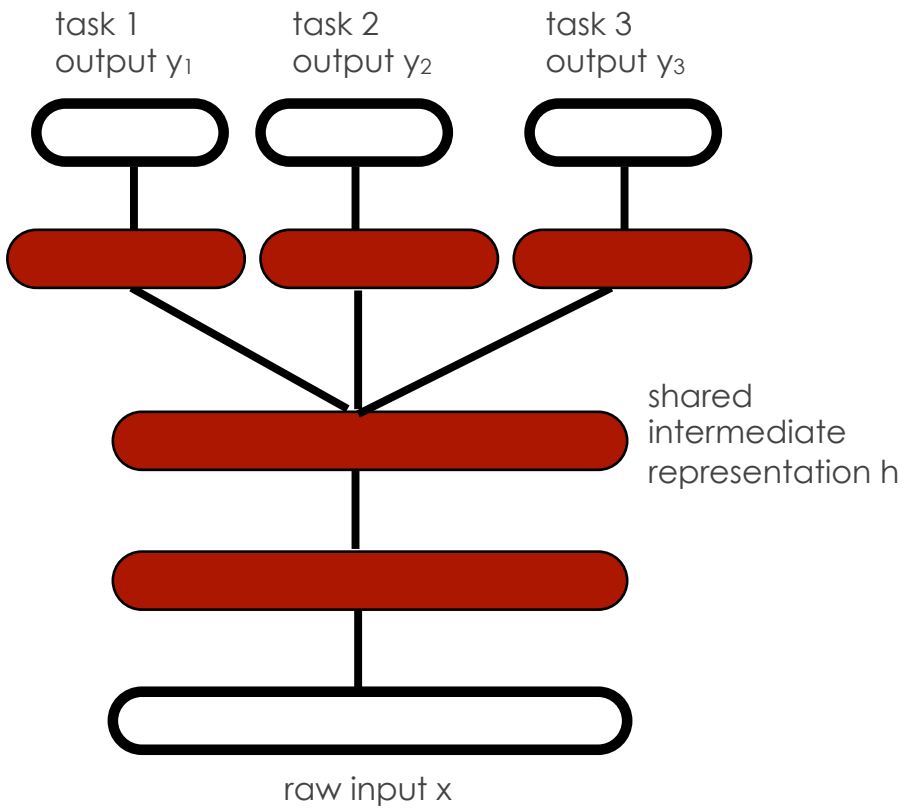
Theorems for all 3:
(Hastad et al 86 & 91, Bengio et al 2007)

Functions compactly represented with k layers may require exponential size with k-1 layers



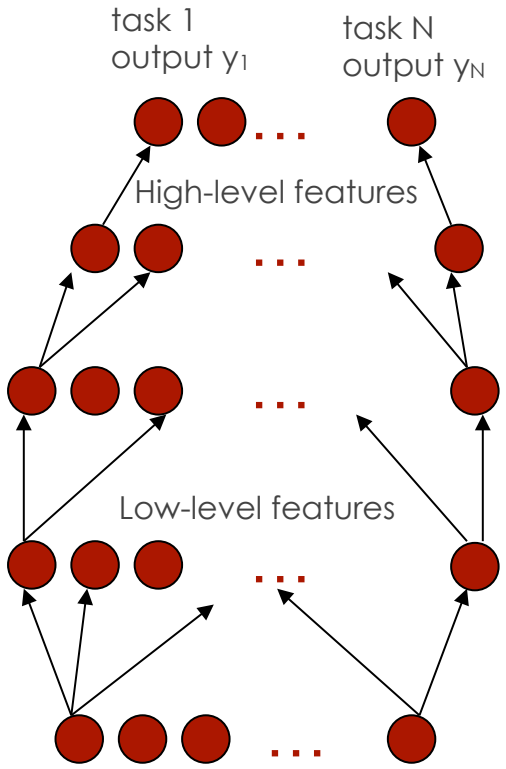
Deep Architectures and Sharing Statistical Strength, Multi-Task Learning

- Generalizing better to new tasks is crucial to approach AI
- Deep architectures learn good intermediate representations that can be shared across tasks
- A good representation is one that makes sense for many tasks

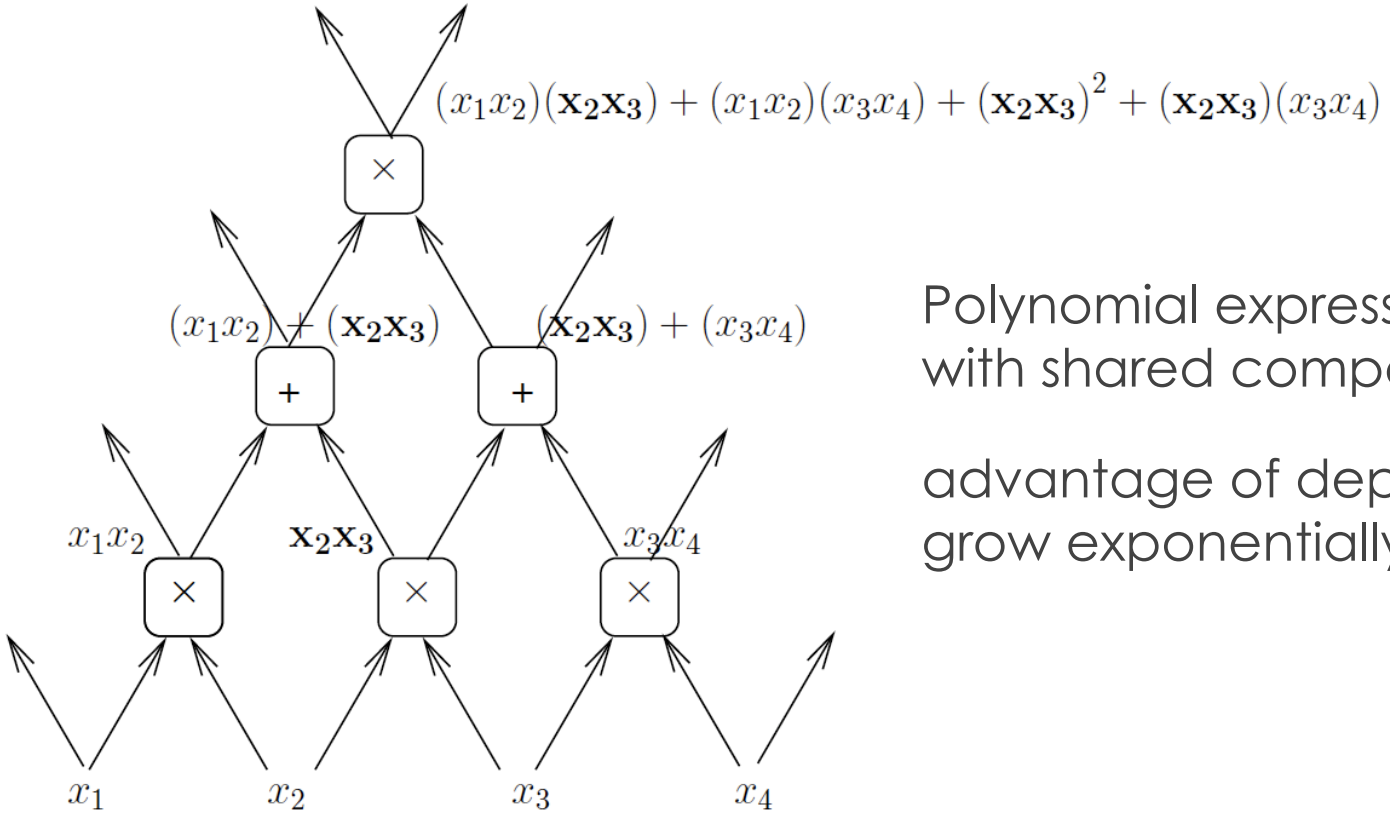


Feature and Sub-Feature Sharing

- Different tasks can share the same high-level feature
- Different high-level features can be built from the same set of lower-level features
- More levels = up to exponential gain in representational efficiency



Sharing Components in a Deep Architecture



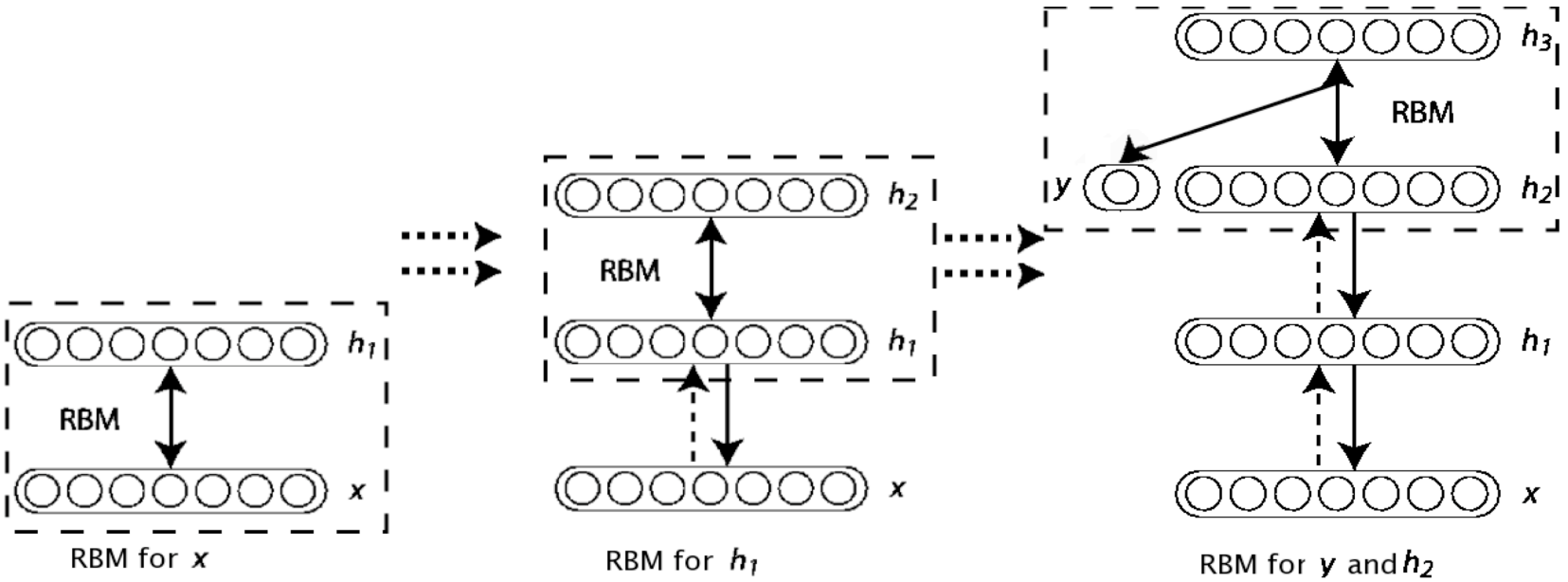
Polynomial expressed with shared components:

advantage of depth may grow exponentially

The Deep Breakthrough

- Before 2006, training deep architectures was unsuccessful, except for convolutional neural nets
- Hinton, Osindero & Teh « A Fast Learning Algorithm for Deep Belief Nets », *Neural Computation*, 2006
- Bengio, Lamblin, Popovici, Larochelle « Greedy Layer-Wise Training of Deep Networks », *NIPS'2006*
- Ranzato, Poultney, Chopra, LeCun « Efficient Learning of Sparse Representations with an Energy-Based Model », *NIPS'2006*

Greedy Layer-Wise Pre-Training



Stacking Restricted Boltzmann Machines (RBM) → Deep Belief Network (DBN)

Greedy Layer-Wise Unsupervised Pre-Training Algorithm

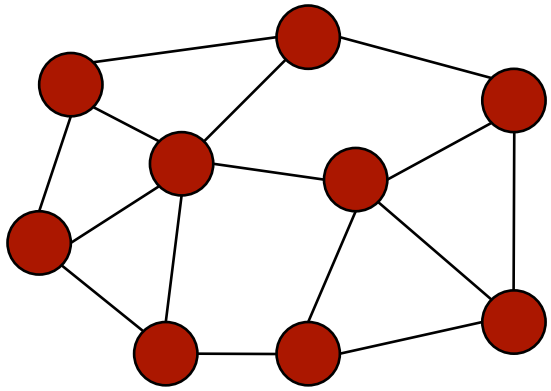
- Train unsupervised feature extractor (e.g. RBM, auto-encoder) mapping input x to representation h_1, capturing main factors of variation in x (models $P(x)$)
- Taking $h_1(x)$ as an input, train a second unsupervised feature extractor, obtaining representation h_2 of x
- Etc. to level k gives h_k
- Plug a supervised classifier $P(Y \mid h_k(x))$ on top, taking h_k as input
- Fine-tune parameters of whole system $P(Y \mid x)$ wrt supervised objective

Boltzman Machines and MRFs

- Boltzmann machines: $P(x) = \frac{1}{Z} e^{-\text{Energy}(x)} = e^{c^T x + x^T W x}$
  (Hinton 84)

- Markov Random Fields:

$$P(x) = \frac{1}{Z} e^{\sum_i w_i f_i(x)}$$

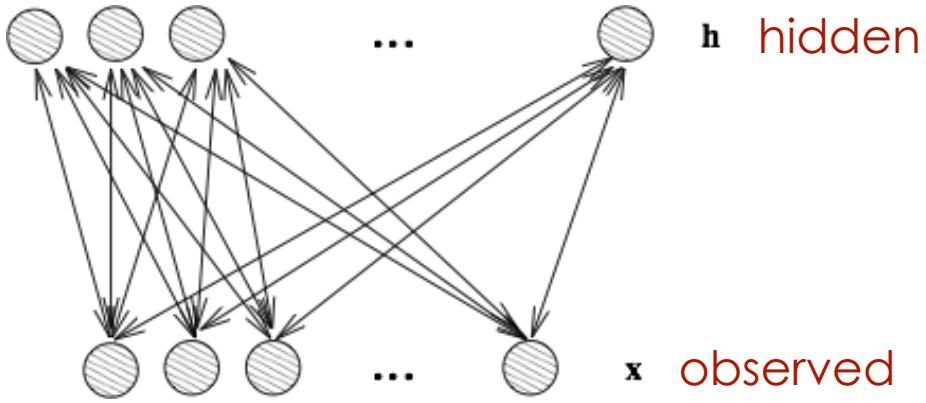


- More interesting with latent variables!

Restricted Boltzman Machine

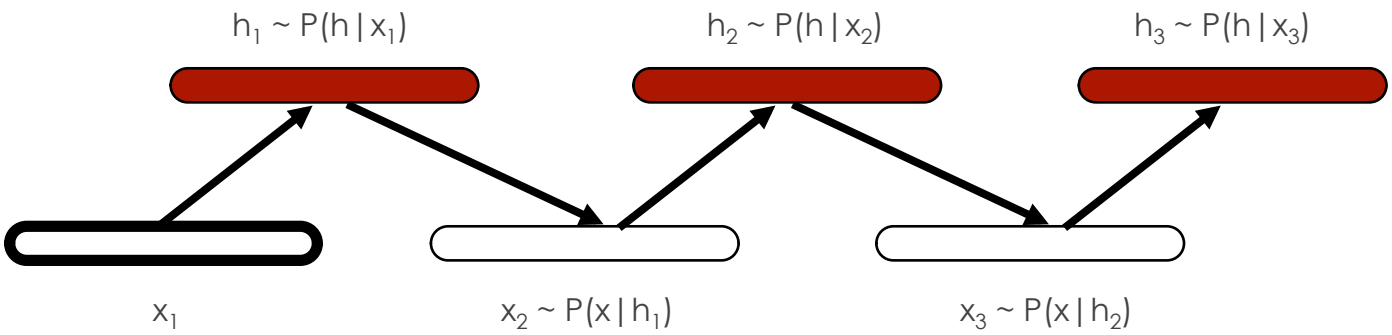
- The most popular building block for deep architectures
  (Smolensky 86, Hinton 2002)

$$P(x,h) = \frac{1}{Z} e^{b^T h + c^T x + h^T W x}$$



- **Bipartite** undirected graphical model
- *h~P(h | x)*, or *(P(h$_i$=1 | x))* are representations of x

Gibbs Sampling in RBMs



P(h | x) and P(x | h) factorize

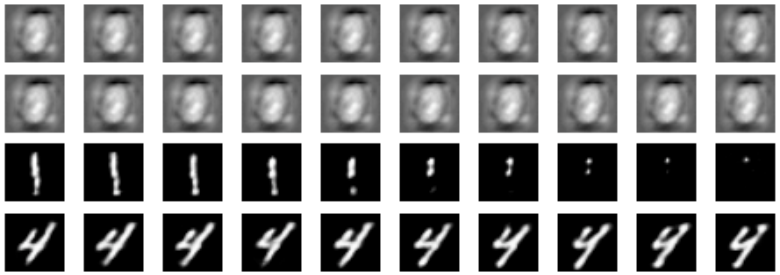
$$P(x,h) = \frac{1}{Z} e^{b^T h + c^T x + h^T W x}$$

- Easy inference
- Convenient Gibbs sampling
  x→h→x→h…

Problems with Gibbs Sampling

In practice, Gibbs sampling does not always mix well…

RBM trained by CD on MNIST



Chains from random state

Chains from real digits

Boltzmann Machine Gradient

$$P(x) = \frac{1}{Z}\sum_h e^{-\text{Energy}(x,h)} = \frac{1}{Z}e^{-\text{FreeEnergy}(x)}$$

- Gradient has two components:

"positive phase" "negative phase"

$$\frac{\partial \log P(x)}{\partial \theta} = -\frac{\partial \text{FreeEnergy}(x)}{\partial \theta} + \sum_{\tilde{x}} P(\tilde{x}) \frac{\partial \text{FreeEnergy}(\tilde{x})}{\partial \theta}$$

$$= -\sum_h P(h|x) \frac{\partial \text{Energy}(x,h)}{\partial \theta} + \sum_{\tilde{x},\tilde{h}} P(\tilde{x},\tilde{h}) \frac{\partial \text{Energy}(\tilde{x},\tilde{h})}{\partial \theta}$$

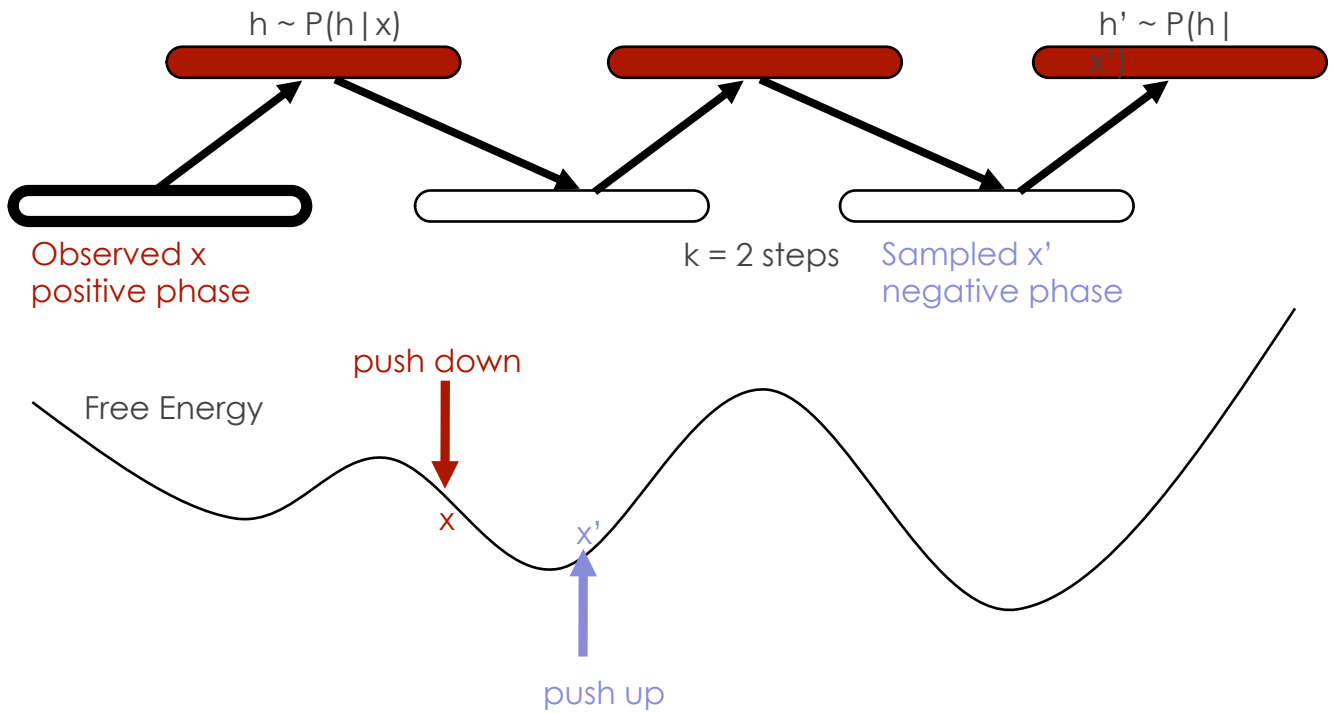
- In RBMs, easy to sample or sum over h | x
- Difficult part: sampling from P(x), typically with a Markov chain

Training RBMs

Contrastive Divergence: (CD-k)	start negative Gibbs chain at observed x, run k Gibbs steps
Persistent CD: (PCD)	run negative Gibbs chain in background while weights slowly change
Fast PCD:	two sets of weights, one with a large learning rate only used for negative phase, quickly exploring modes
Herding:	(see Max Welling's ICML, UAI and ICML workshop talks)
Tempered MCMC:	use higher temperature to escape modes

Contrastive Divergence

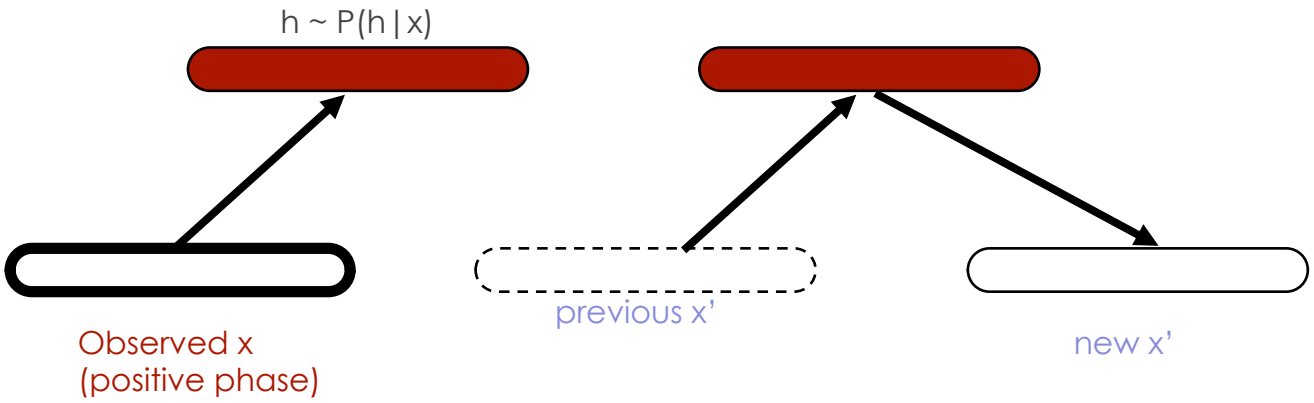
Contrastive Divergence (CD-k): start negative phase block Gibbs chain at observed x, run k Gibbs steps (Hinton 2002)



Persistent CD (PCD)

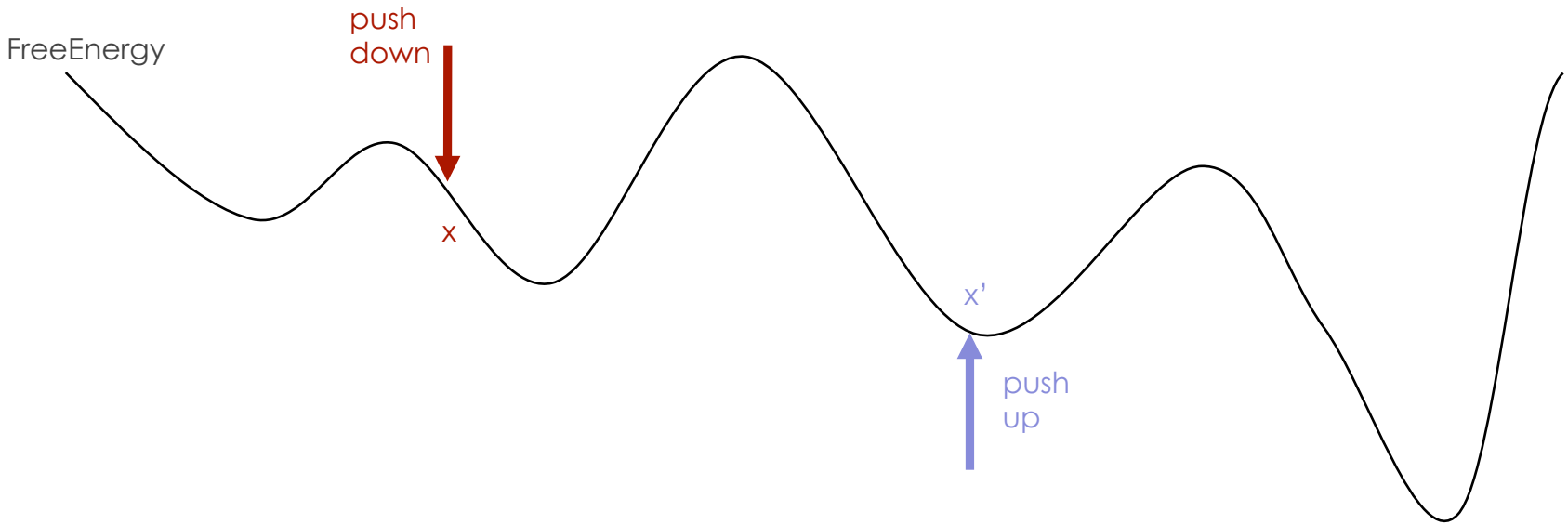
Run negative Gibbs chain in background while weights slowly change (Younes 2000, Tieleman 2008):

- Guarantees (Younes 89, 2000; Yuille 2004)
- If learning rate decreases in 1/t,
  chain mixes before parameters change too much,
  chain stays converged when parameters change



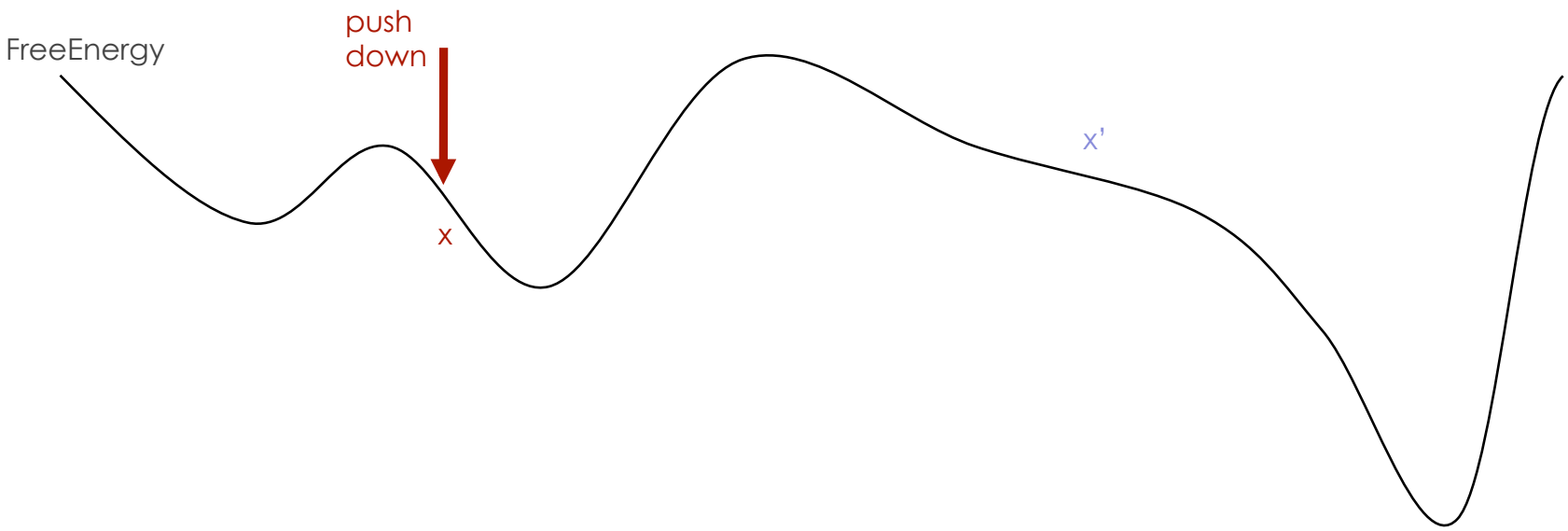
Persistent CD with Large Step Size

Negative phase samples quickly push up the energy of wherever they are and quickly move to another mode



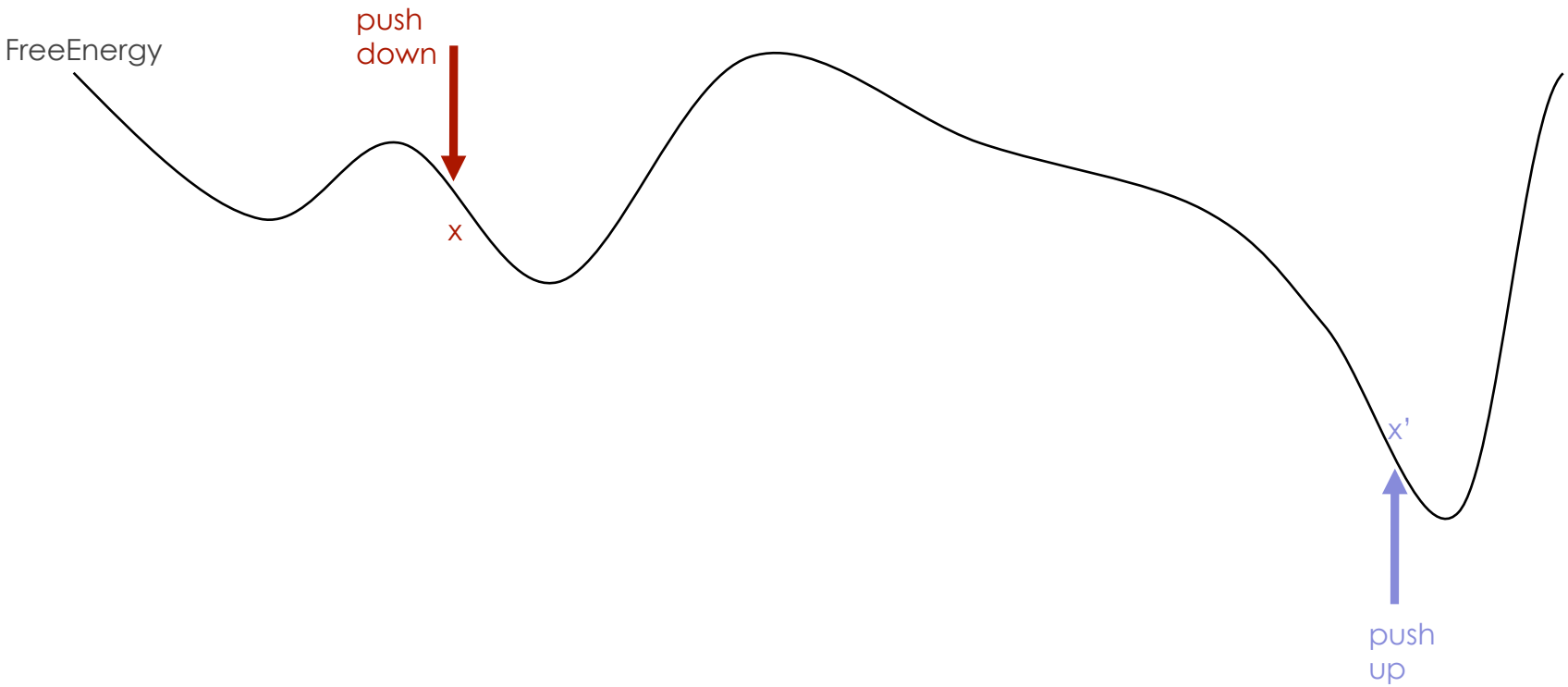
Persistent CD with Large Step Size

Negative phase samples quickly push up the energy of wherever they are and quickly move to another mode



Persistent CD with Large Step Size

Negative phase samples quickly push up the energy of wherever they are and quickly move to another mode



Fast Persistent CD and Herding

- Exploit impressively faster mixing achieved when parameters change quickly (large learning rate) while sampling
- Fast PCD: two sets of weights, one with a large learning rate only used for negative phase, quickly exploring modes
- Herding (see Max Welling's ICML, UAI and ICML workshop talks): 0-temperature MRFs and RBMs, only use fast weights

Herding MRFs

(Welling, ICML'2009)

- Consider 0-temperature MRF with state *s* and weights *w*

$$\ell = E_{data\ s^+}[\sum_i w_i f_i(s^+)] - \max_s \sum_i w_i f_i(s)$$

- Fully observed case, observe values s^+, dynamical system where s^- and *W* evolve

$$s^- \leftarrow \text{argmax}_s \sum_i w_i f_i(s)$$
$$w \leftarrow w + E_{data\ s^+}[f(s^+)] - f(s^-)$$

- Then statistics of samples s^- match the data's statistics, even if approximate max, as long as *w* remains bounded

$$E_{samples\ s^-}[f(s^-)] = E_{data\ s^+}[f(s^+)]$$

Herding RBMs

(Welling, UAI'2009)

- Hidden part *h* of the state $s = (x,h)$
- Binomial state variables $s_i \in \{-1,1\}$
- Statistics *f* s_i, $s_i\, s_j$
- Optimize *h* given x in positive phase

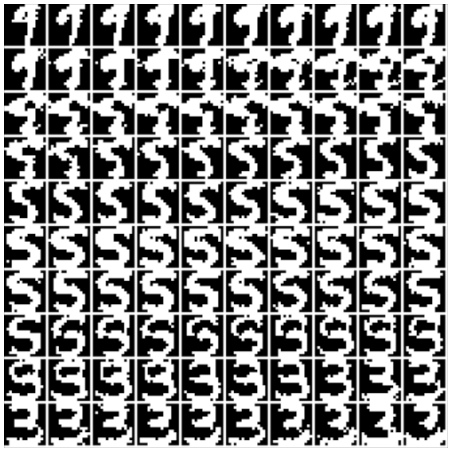
$$\ell = E_{data\, x^+}[\max_h \sum_i w_i f_i(x^+,h)] - \max_s \sum_i w_i f_i(s)$$

$$s^- \leftarrow \mathrm{argmax}_s \sum_i w_i f_i(s)$$

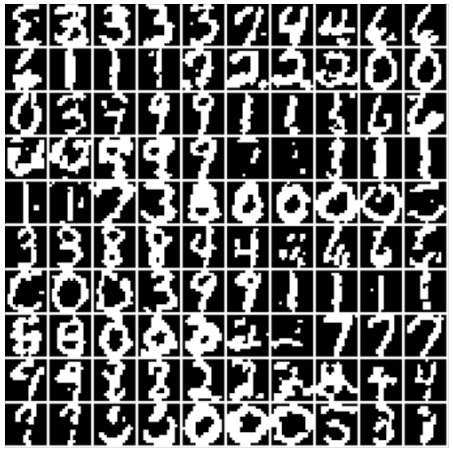
$$w \leftarrow w + E_{data\, x^+}[\max_h f(x^+,h)] - f(s^-)$$

- In practice, greedy maximization works, exploiting RBM structure. Works like 0-temperature PCD.

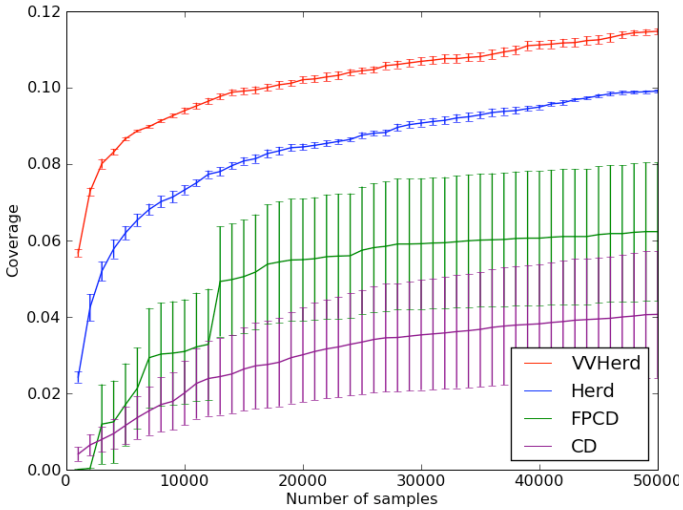
Fast Mixing with Herding



FPCD



Herding



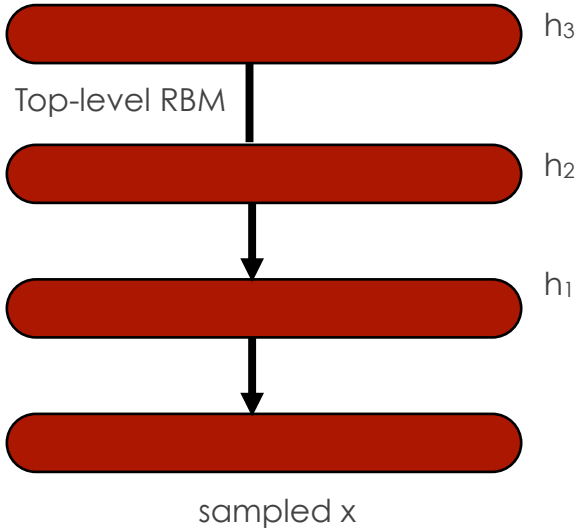
Tempered MCMC

- Annealing from high-temperature worked well for estimating log-likelihood (AIS)
- Consider multiple chains at different temperatures and reversible swaps between adjacent chains
- Higher temperature chains can escape modes
- Model samples are from T=1

	Sample Generation Procedure		
Training Procedure	TMCMC	Gibbs (ramdom start)	Gibbs (test start)
TMCMC	215.4 ± 2.2	88.4 ± 2.7	60.0 ± 2.8
PCD	44.7 ± 2.5	-28.6 ± 3.2	-175.1 ± 2.9
CD	-2165 ± 0.5	-2154 ± 0.6	-842.8 ± 6.1

Deep Belief Networks

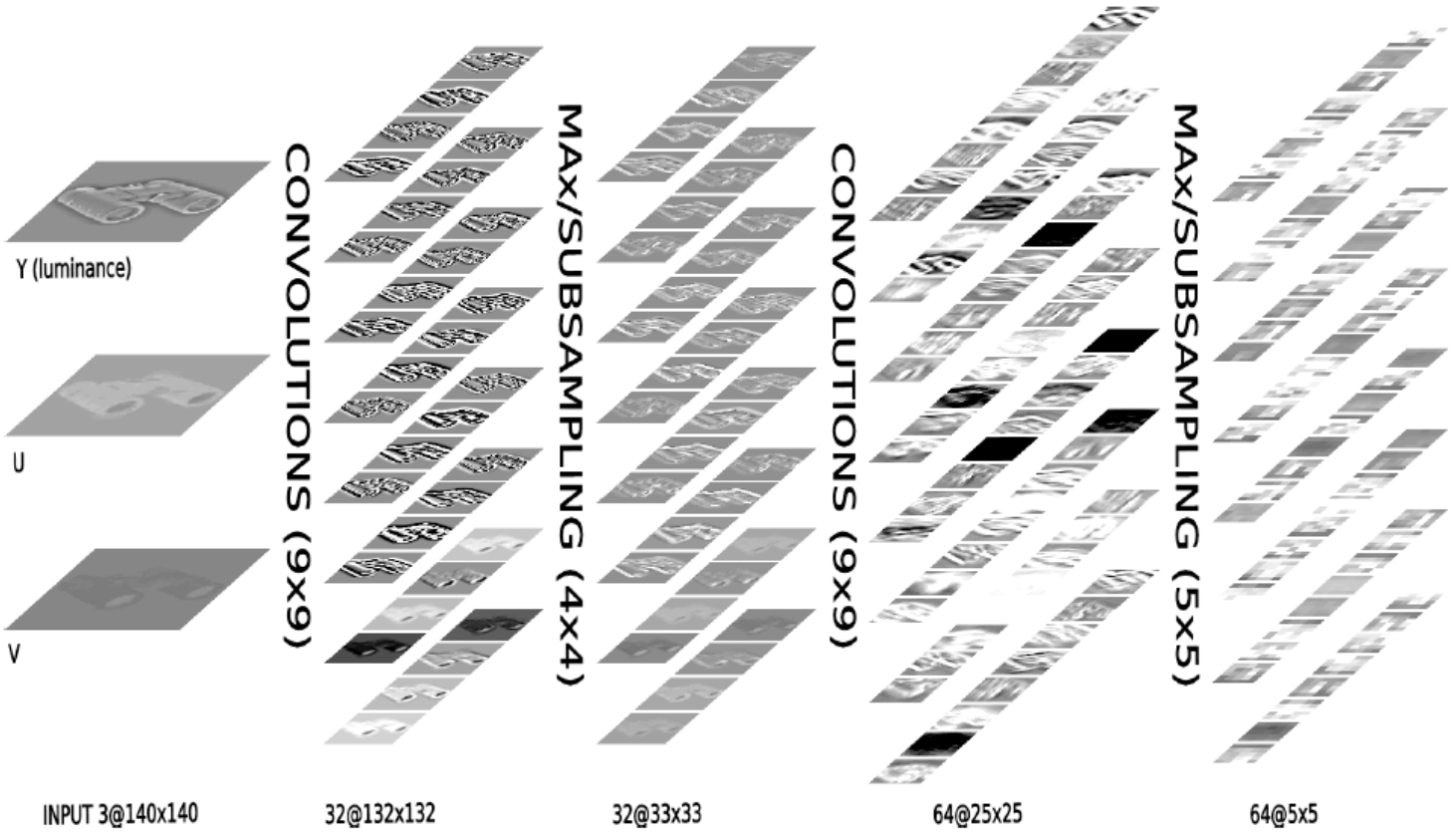
- Sampling:
  - Sample from top RMB
  - Sample from level k given k+1
- Easy approximate inference
- Training:
  - Variational bound justifies greedy layerwise training of RBMs
  - How to train all levels together?



$$\log P(\mathbf{x}) \geq H_{Q(\mathbf{h}|\mathbf{x})} + \sum_{\mathbf{h}} Q(\mathbf{h}|\mathbf{x}) \left(\log P(\mathbf{h}) + \log P(\mathbf{x}|\mathbf{h})\right)$$

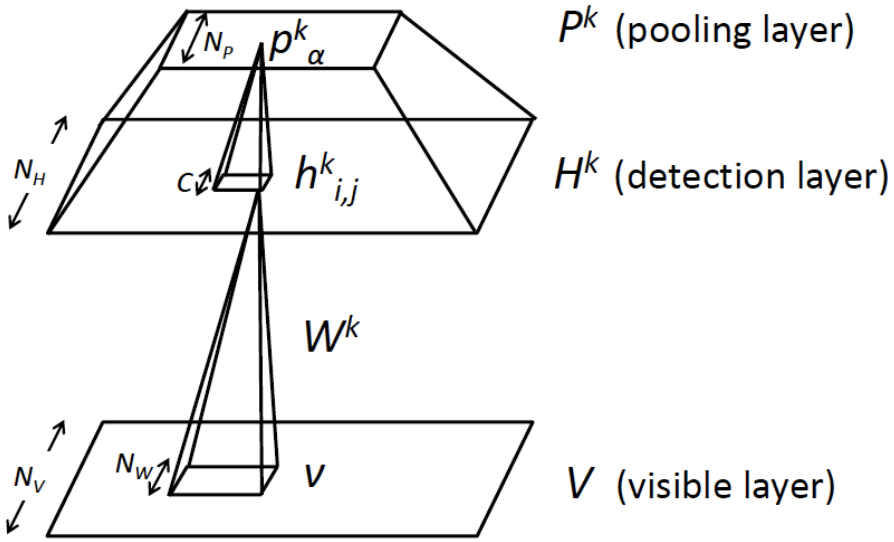
Deep Convolutional Architectures

Mostly from Le Cun's (NYU) and Ng's (Stanford) groups:
state-of-the-art on MNIST digits, Caltech-101 objects, faces



Convolutional DBNs

(Lee et al, ICML'2009)

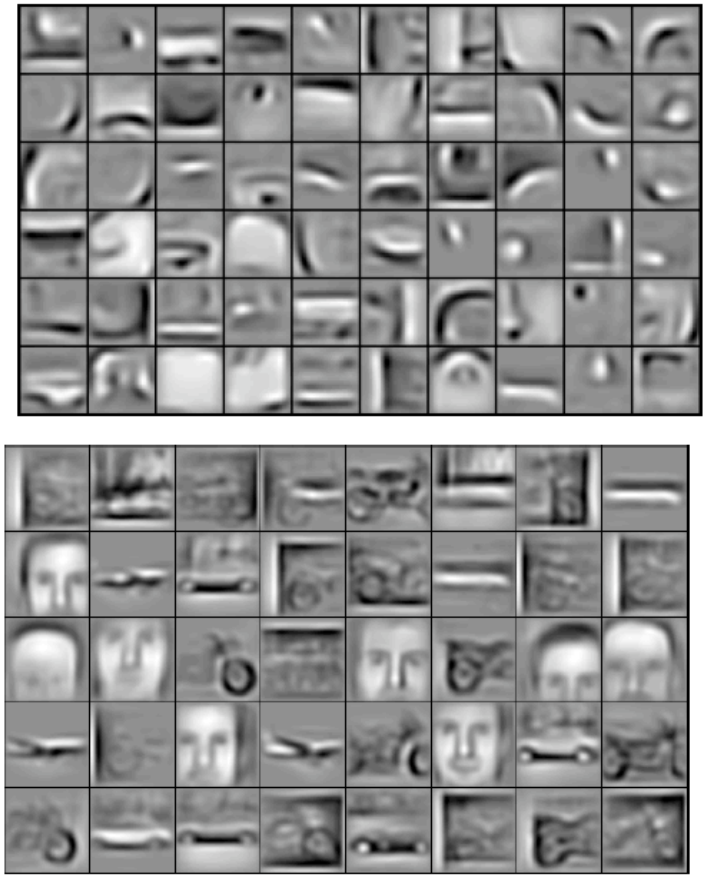


P^k (pooling layer)

H^k (detection layer)

V (visible layer)

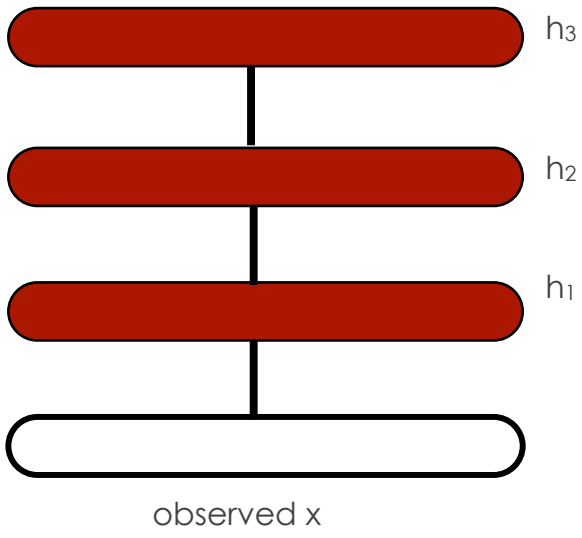
faces, cars, airplanes, motorbikes



Deep Boltzman Machines

(Salakhutdinov et al, AISTATS 2009, Lee et al, ICML 2009)

- Positive phase: variational approximation (mean-field)
- Negative phase: persistent chain
- **Can (must) initialize from stacked RBMs**
- Improved performance on MNIST from 1.2% to .95% error
- Can apply AIS with 2 hidden layers



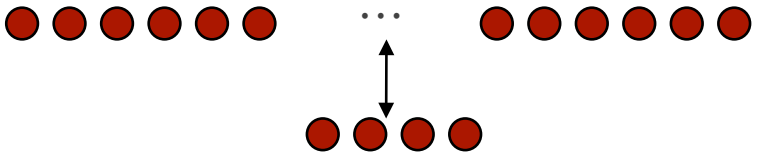
Estimating Log-Likelihood

- RBMs: requires estimating partition function
  - Reconstruction error provides a cheap proxy
  - Log Z tractable analytically for < 25 binary inputs or hidden
  - Lower-bounded (how well?) with Annealed Importance Sampling (AIS)
- Deep Belief Networks:

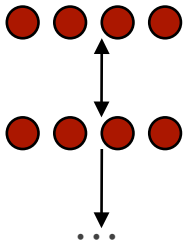
  Extensions of AIS (Salakhutdinov & Murray, ICML 2008, NIPS 2008)
- Open question: efficient ways to monitor progress?

RBMs and Thin DBNs as Non-Parametric Models

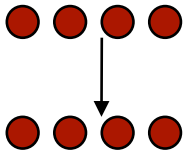
- As the nb of hidden units increases, RBM capacity increases; with enough hidden units (up to 2^d), can approximate any discrete distribution on d bits (Le Roux & Bengio, 2008)



- Similarly, a thin DBN with d+1 units / layer and enough layers (up to $1+2^d/d$) can approximate any discrete distribution.



- Same proof technique also shows universal approximation property of thin deep deterministic neural networks



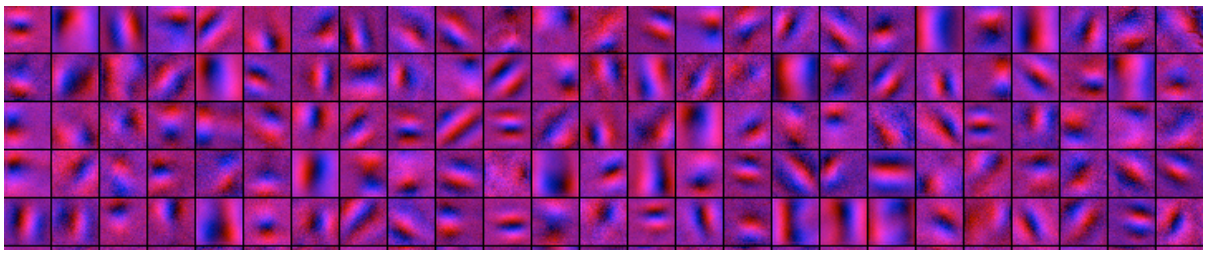
(Le Roux & Bengio, 2009, submitted)

Why are Classifiers Obtained from DBNs Working so Well?

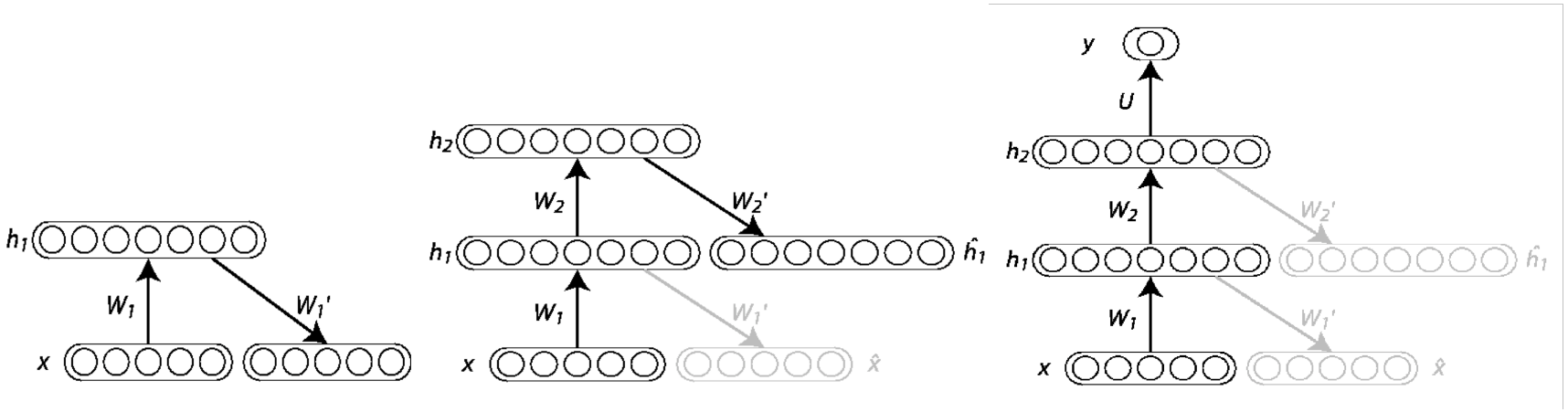
- General principles?
- Would these principles work for other single-level algorithms?
- Why does it work?

Replacing RBMs by Other Layer-Local Unsupervised Learning

- Auto-encoders (Bengio et al, NIPS'2006)
- Sparse auto-encoders (Ranzato et al, NIPS'2006)
- Kernel PCA (Erhan 2008)
- Denoising auto-encoders (Vincent et al, ICML'2008)
- Unsupervised embedding (Weston et al, ICML'2008)
- Slow features (Mohabi et al, ICML'2009, Bergstra & Bengio NIPS'2009)



Stacking Auto-Encoders



Auto-Encoders and CD

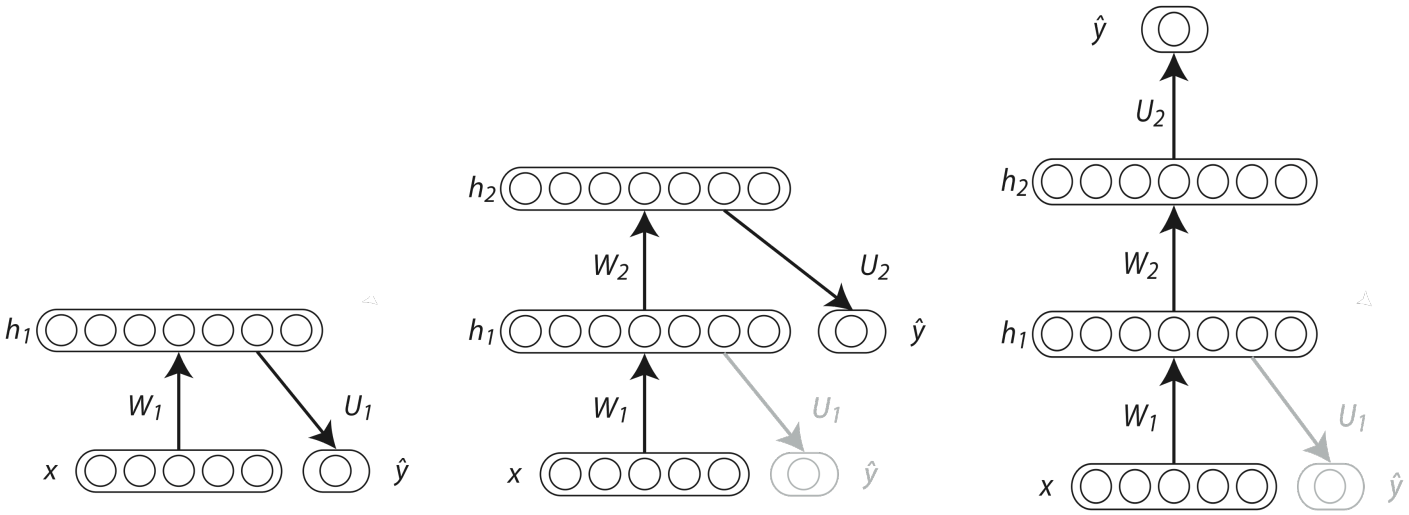
RBM log-likelihood gradient written as converging expansion:

- CD-k = 2 k terms
- reconstruction error ~ 1 term

$$\frac{\partial \log P(x_1)}{\partial \theta} = \sum_{s=1}^{t-1} \left(E\left[\frac{\partial \log P(x_s|h_s)}{\partial \theta} \middle| x_1 \right] + E\left[\frac{\partial \log P(h_s|x_{s+1})}{\partial \theta} \middle| x_1 \right] \right) + E\left[\frac{\partial \log P(x_t)}{\partial \theta} \middle| x_1 \right]$$

(Bengio & Delalleau 2009)

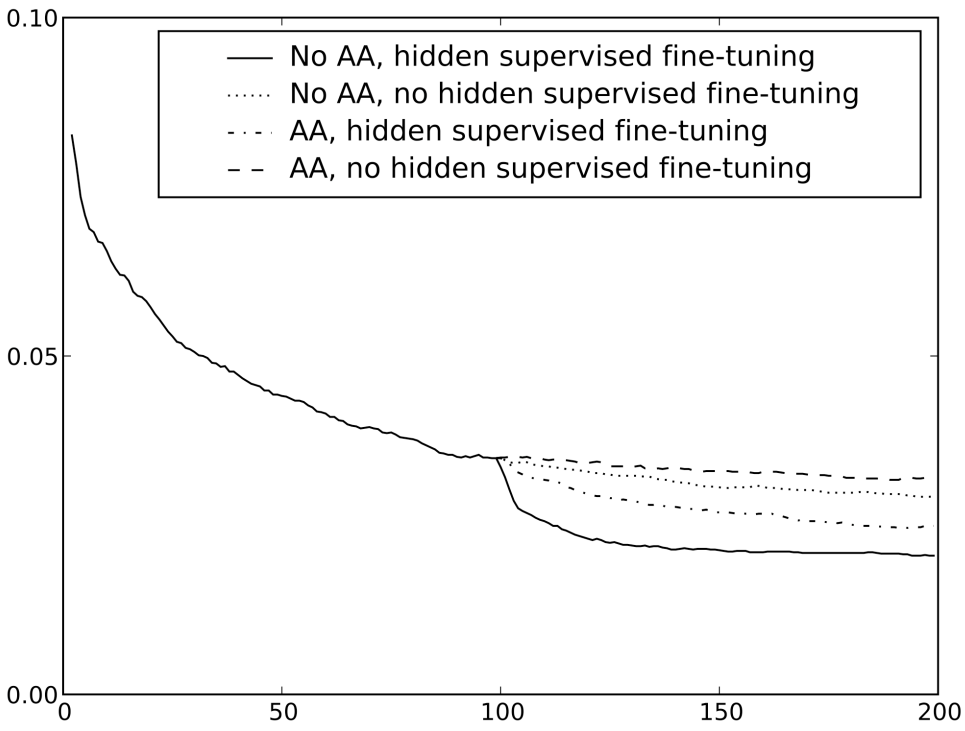
Greedy Layerwise Supervised Training



Generally worse than unsupervised pre-training but better than ordinary training of a deep neural network (Bengio et al. 2007).

Supervised Fine-Tuning is Important

- Greedy layer-wise unsupervised pre-training phase with RBMs or auto-encoders on MNIST
- Supervised phase with or without unsupervised updates, with or without fine-tuning of hidden layers
- Can train all RBMs at the same time, same results



Sparse Auto-Encoders

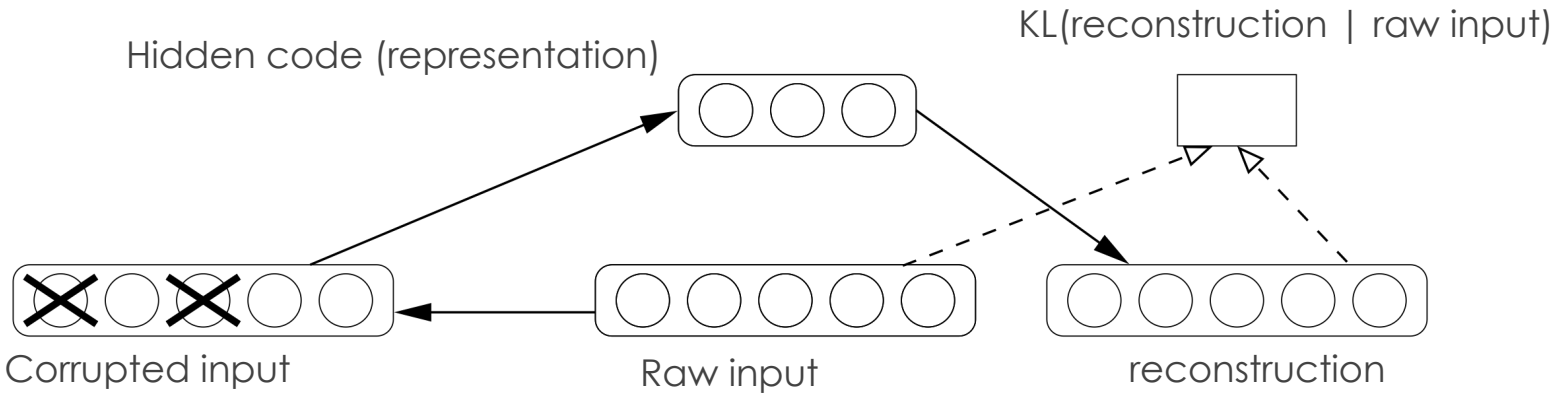
(Ranzato et al, 2007; Ranzato et al 2008)

- Sparsity penalty on the intermediate codes
- Like sparse coding and has explaining away but with efficient run-time encoder (approximate posterior)
- Sparsity penalty pushes up the free energy of all configurations (proxy for minimizing the partition function)
- Impressive results in object classification (convolutional nets):
  - MNIST 0.5% error = record-breaking
  - Caltech-101 65% correct = state-of-the-art (Jarrett et al, ICCV 2009)
- Similar results obtained with a convolutional DBN (Lee et al, ICML'2009)

Denoising Auto-Encoder

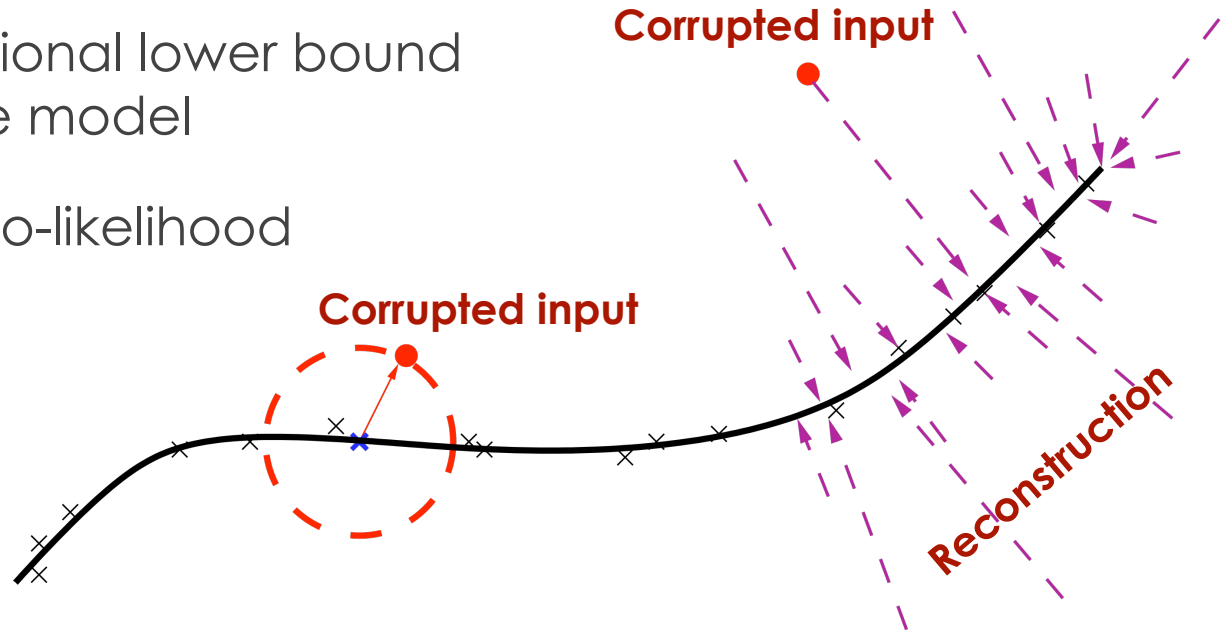
(Vincent et al, ICML 2008)

- Corrupt the input
- Reconstruct the uncorrupted input



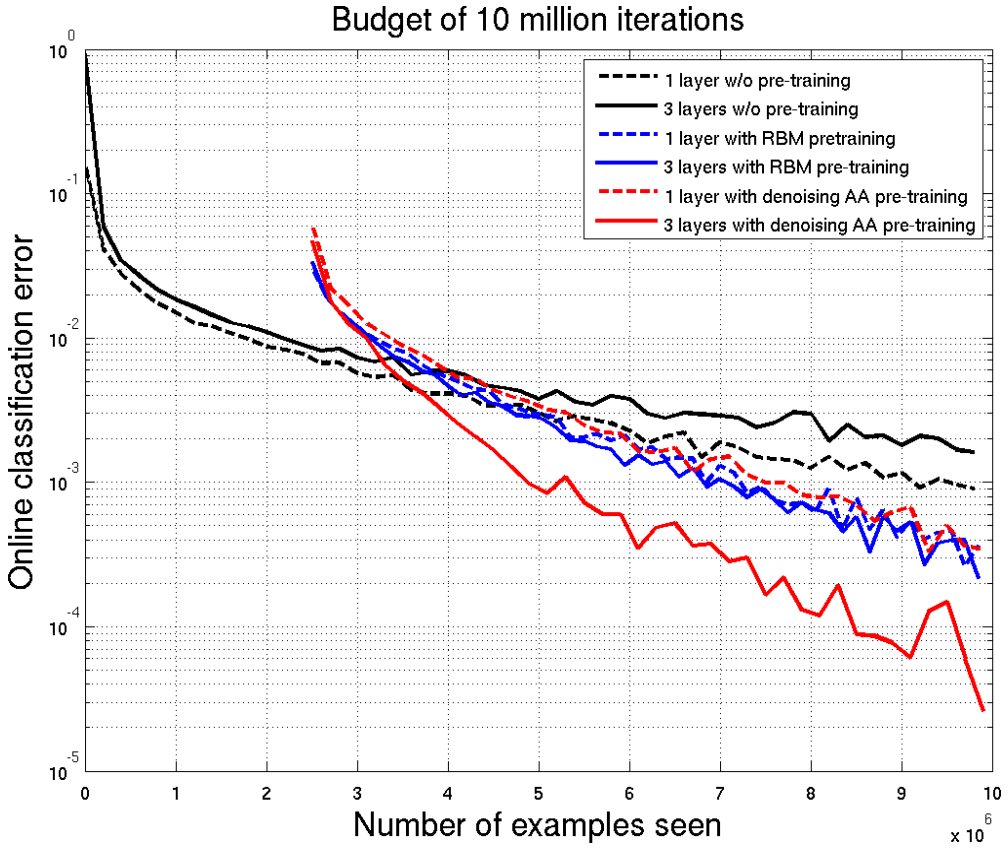
Denoising Auto-Encoder

- Learns a vector field towards higher probability regions
- Minimizes variational lower bound on a generative model
- Similar to pseudo-likelihood



Stacked Denoising Auto-Encoders

- No partition function, can measure training criterion
- Encoder & decoder: any parametrization
- Performs as well or better than stacking RBMs for usupervised pre-training



Infinite MNIST

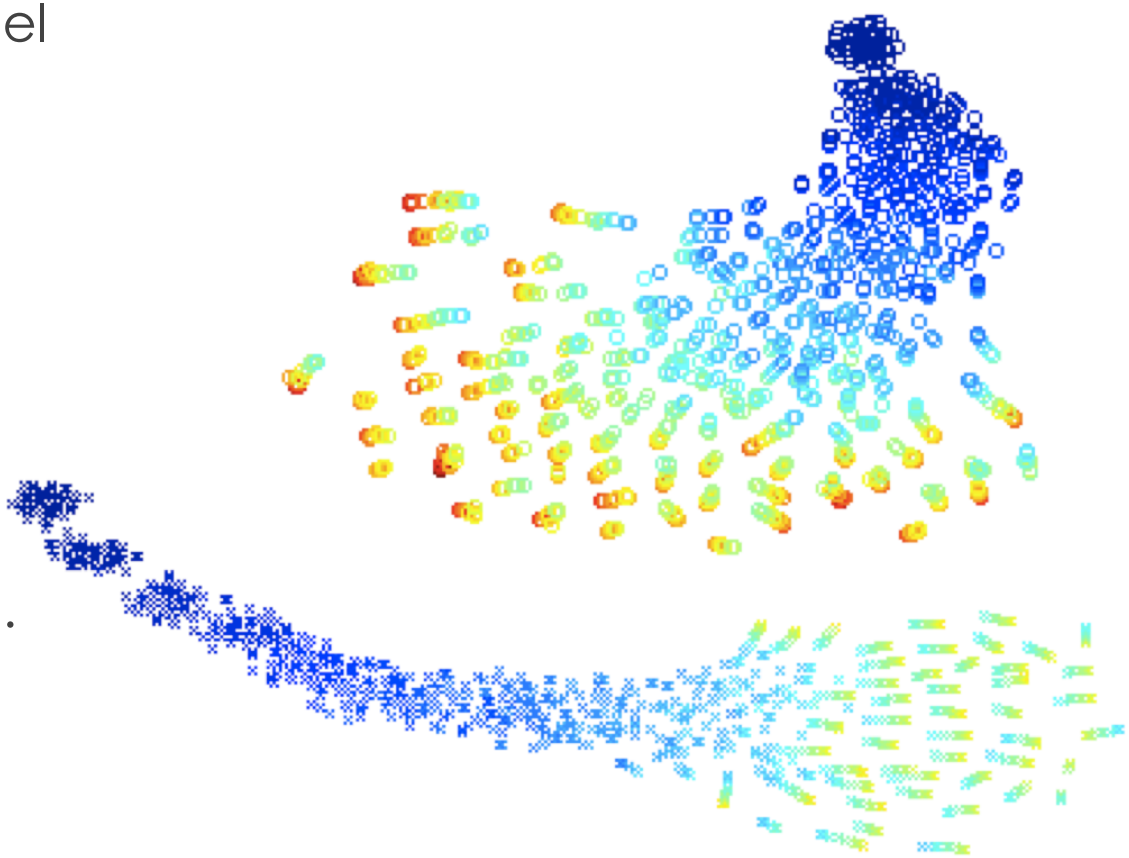
Why is Unsupervised Pre-Training Working So Well?

- Regularization hypothesis:
  - Unsupervised component forces model close to P(x)
  - Representations good for P(x) are good for P(y | x)

- Optimization hypothesis:
  - Unsupervised initialization near better local minimum of P(y | x)
  - Can reach lower local minimum otherwise not achievable by random initialization

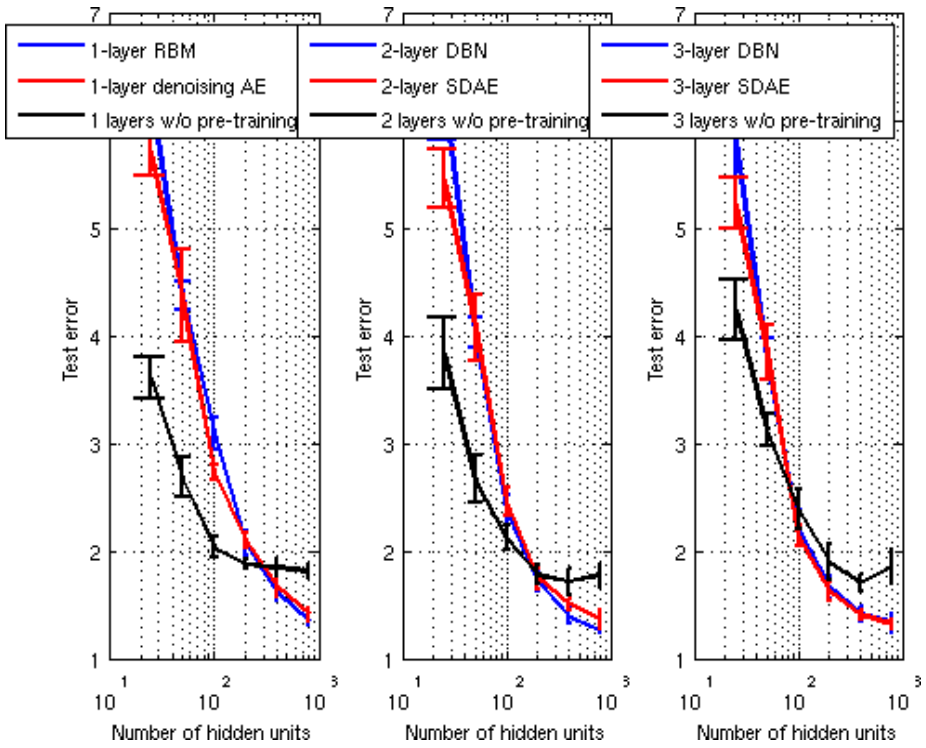
Learning Trajectories in Function Space

- Each point a model in function space
- Color = epoch
- Top: trajectories w/o pre-training
- Each trajectory converges in different local min.
- No overlap of regions with and w/o pre-training



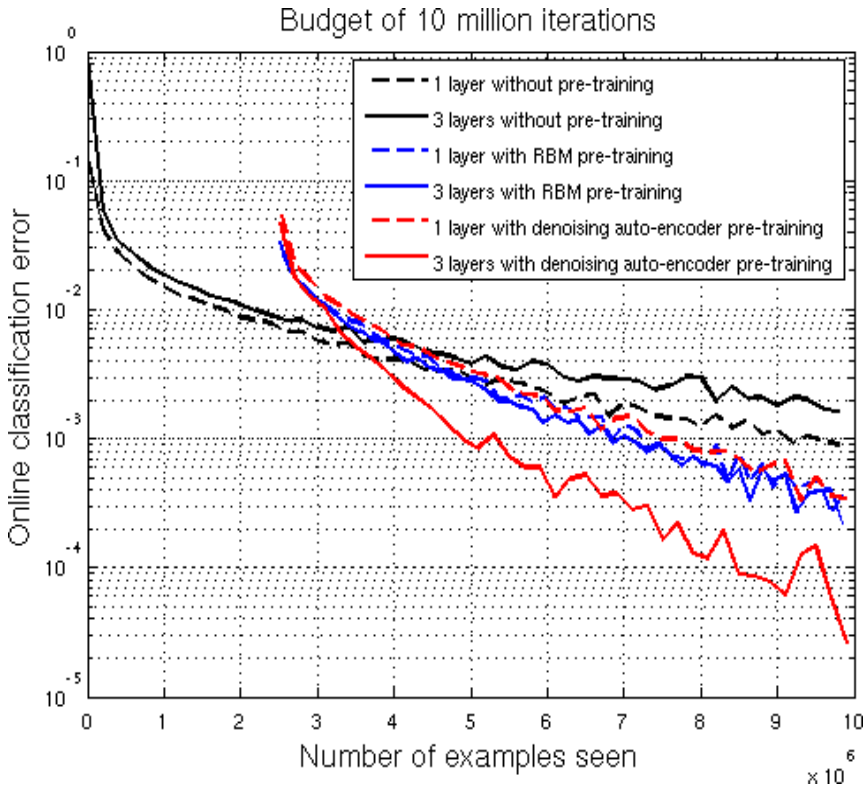
Unsupervised Learning as Regularizer

- Adding extra regularization (reducing # hidden units) hurts more the pre-trained models
- Pre-trained models have less variance wrt training sample
- Regularizer = infinite penalty outside of region compatible with unsupervised pre-training

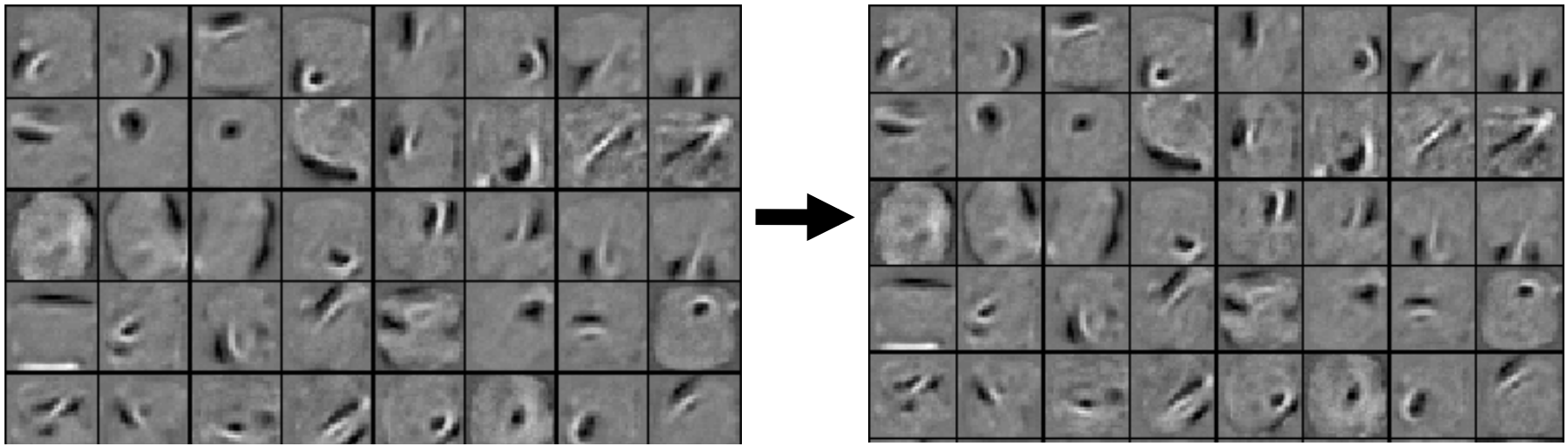


Better Optimization of Online Error

- Both training and online error are smaller with unsupervised pre-training
- As # samples → ∞ training err. = online err. = generalization err.
- Without unsup. pre-training: can't exploit capacity to capture complexity in target function from training data



Learning Dynamics of Deep Nets

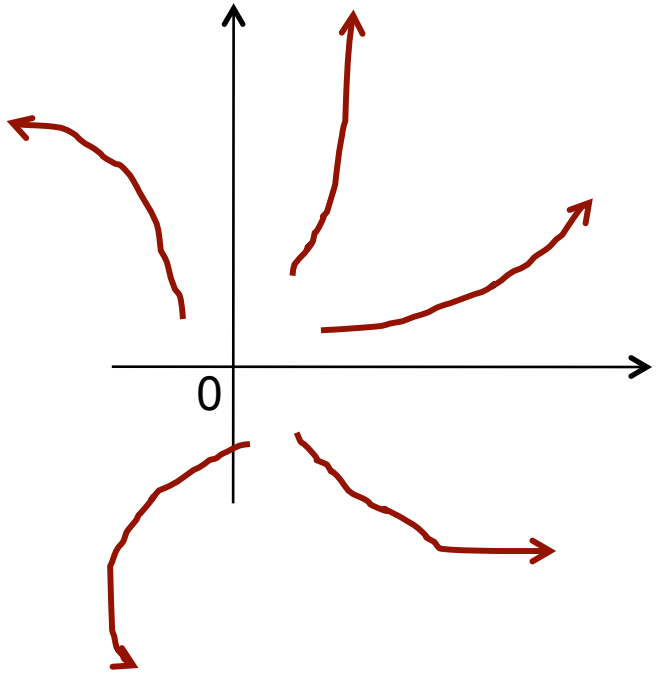


Before fine-tuning

After fine-tuning

Learning Dynamics of Deep Nets

- As weights become larger, get trapped in basin of attraction ("quadrant" does not change)
- Initial updates have a crucial influence ("critical period"), explain more of the variance
- Unsupervised pre-training initializes in basin of attraction with good generalization properties

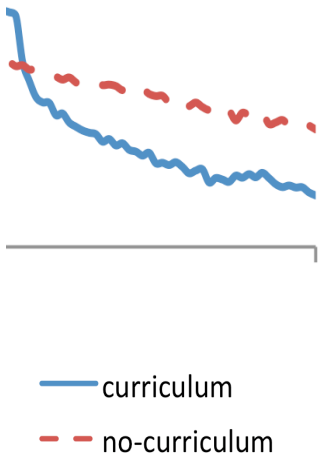


Level-Local Learning is Important

- Initializing each layer of an unsupervised deep Boltzmann machine helps a lot
- Initializing each layer of a supervised neural network as an RBM or auto-encoder helps a lot
- Helps most the layers further away from the target
- Jointly training all the levels of a deep architecture is difficult

Order & Selection of Examples Matters

- Curriculum learning
  (Bengio et al, ICML'2009; Krueger & Dayan 2009)
- Start with easier examples
- Faster convergence to a better local minimum in deep architectures
- Also acts like a regularizer with optimization effect?
- Influencing learning dynamics can make a big difference



Take-Home Messages

- Multiple levels of latent variables: potentially exponential gain in statistical sharing
- RBMs allow fast inference, stacked RBMs / auto-encoders have fast approximate inference
- Gibbs sampling in RBMs sometimes does not mix well, but sampling and learning can interact in surprisingly useful ways
- Unsupervised pre-training of classifiers acts like a strange regularizer with improved optimization of online error
- At least as important as the model: the inference approximations and the learning dynamics

Some Open Problems

- Why is it difficult to train deep architectures?
- What is important in the learning dynamics?
- How to improve joint training of all layers?
- How to sample better from RBMs and deep generative models?
- Monitoring unsupervised learning quality in deep nets?
- Other ways to guide training of intermediate representations?
- More complex models to handle spatial structure of images, occlusion, temporal structure, etc.

Thank you for your attention!

- Questions?
- Comments?