Deep Gated MRF's

Marc'Aurelio Ranzato

ranzato@google.com

www.cs.toronto.edu/~ranzato

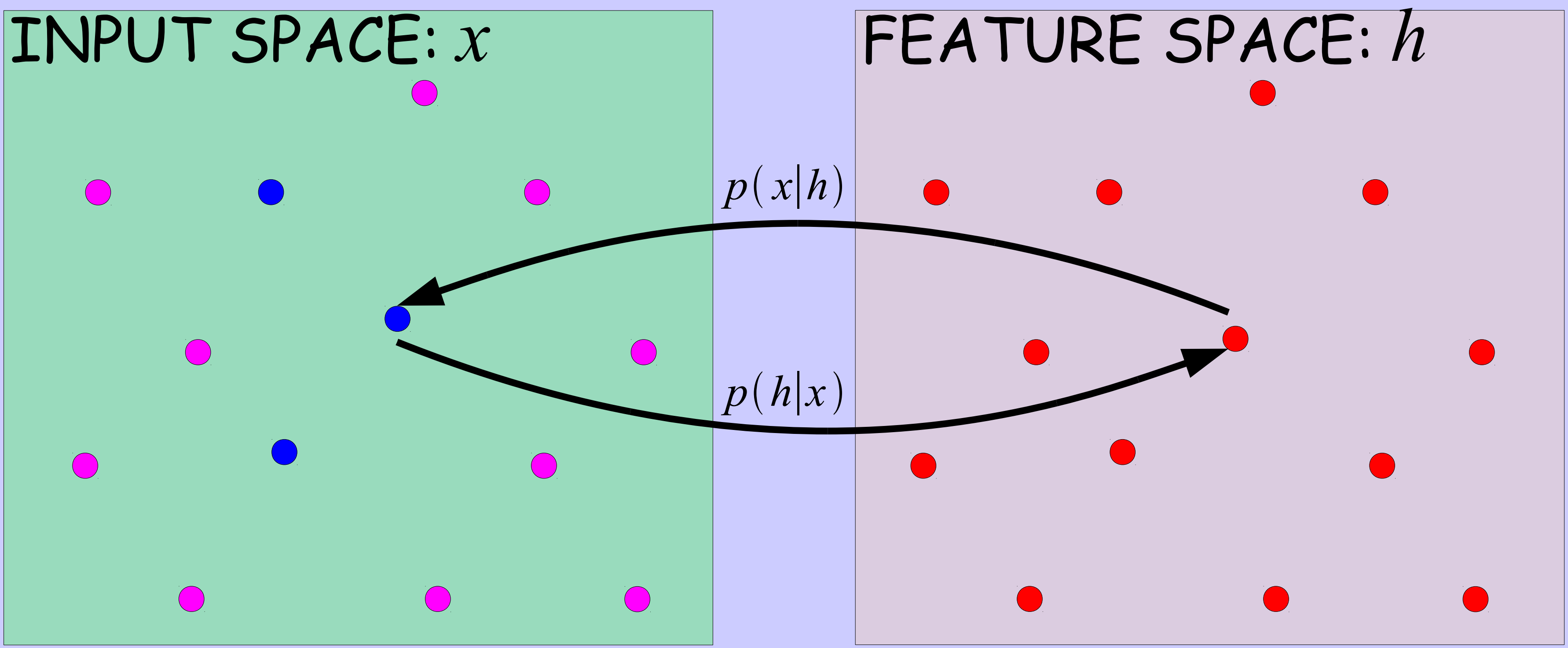
IPAM Summer School, 23 July 2012

Two Approaches to Unsupervised Learning

- structure is learned by scoring input data vectors
- implicit/explicit mapping between input and feature space

Ranzato et al. "A unified energy-based framework for unsupervised learning" AISTATS 2007

- Training sample
- Input vector which is NOT a training sample
- Feature vector

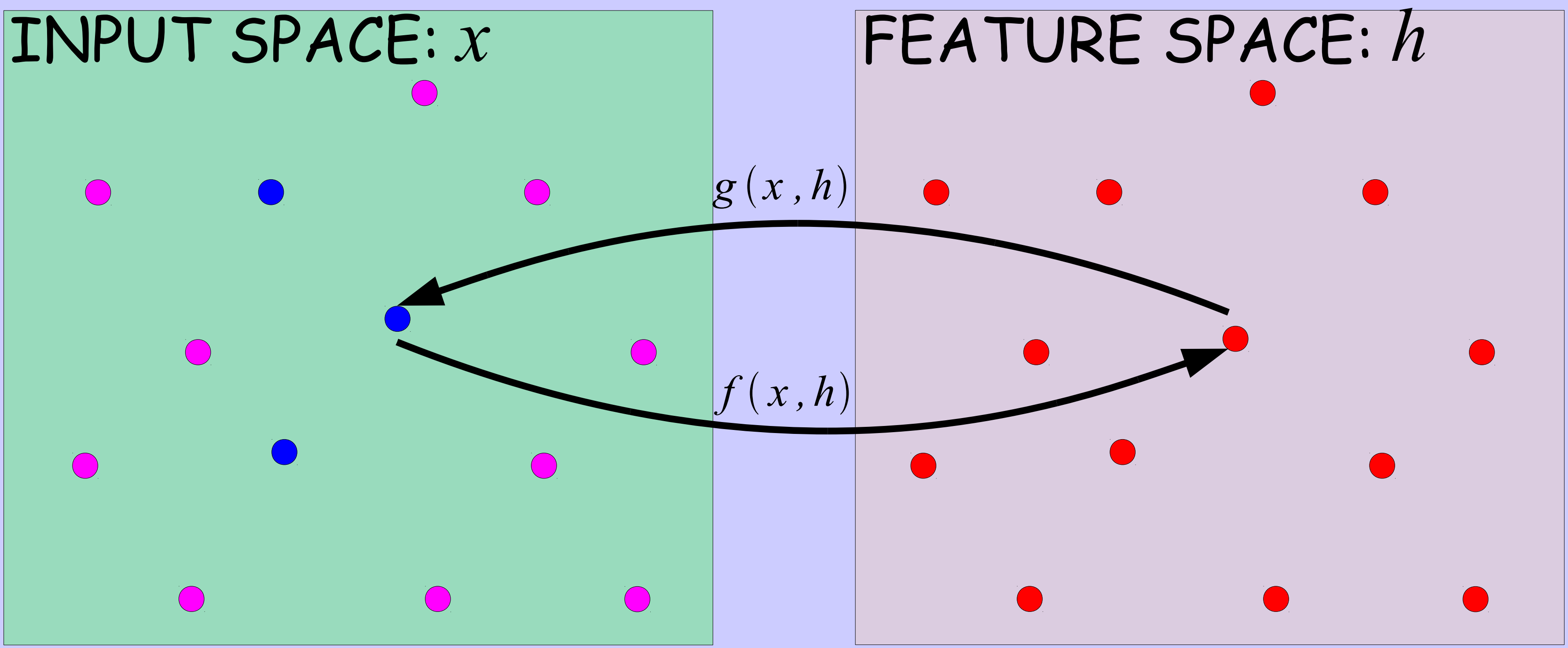


Two Approaches to Unsupervised Learning

- structure is learned by scoring input data vectors
- implicit/explicit mapping between input and feature space

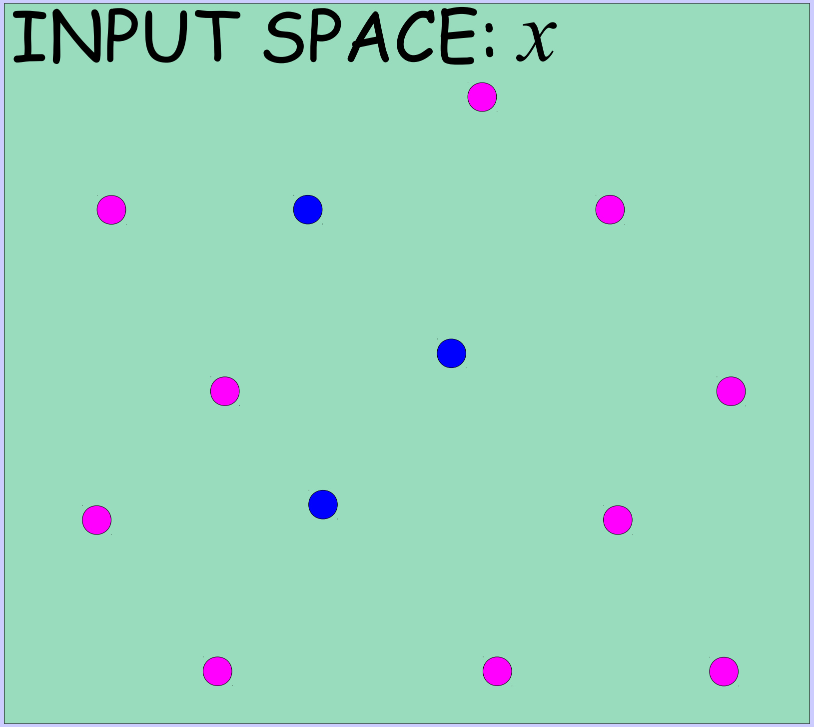
Ranzato et al. "A unified energy-based framework for unsupervised learning" AISTATS 2007

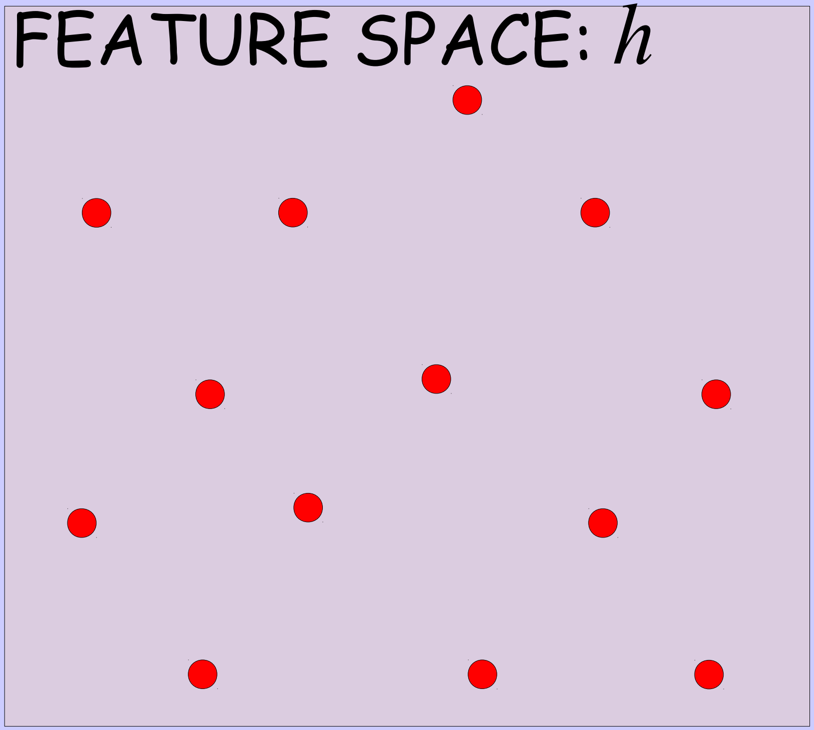
- **Training sample**
- **Input vector which is NOT a training sample**
- **Feature vector**



Two Approaches to Unsupervised Learning

1st strategy: constrain latent representation &
optimize score only at training samples

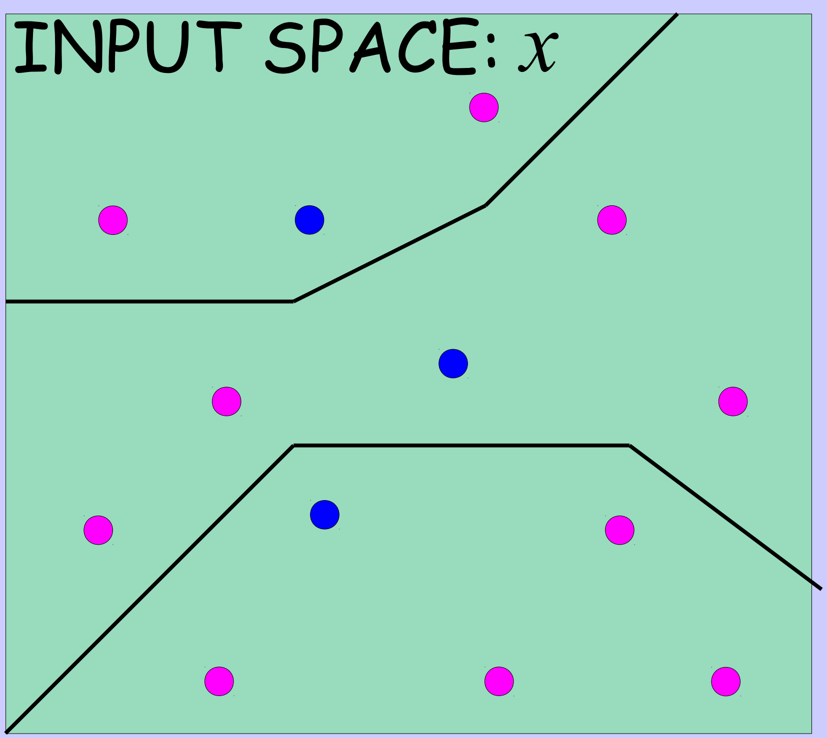


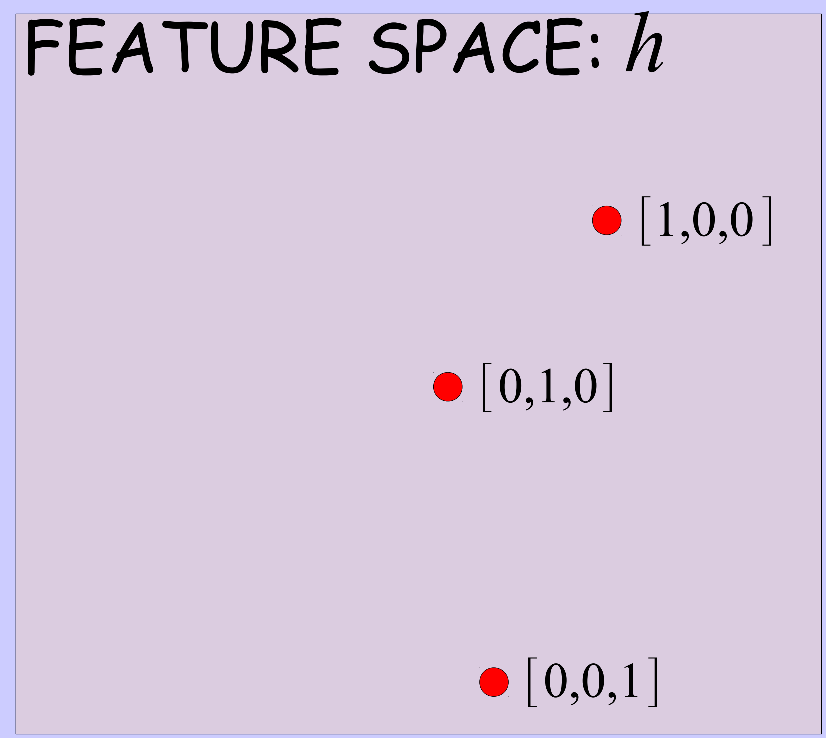


Two Approaches to Unsupervised Learning

1st strategy: constrain latent representation &
optimize score only at training samples

e.g., K-Means: score = reconstruction error: $\|x - Wh\|^2$
constraint = h 1-of-N: [0 0 0 1 0 0 … 0]



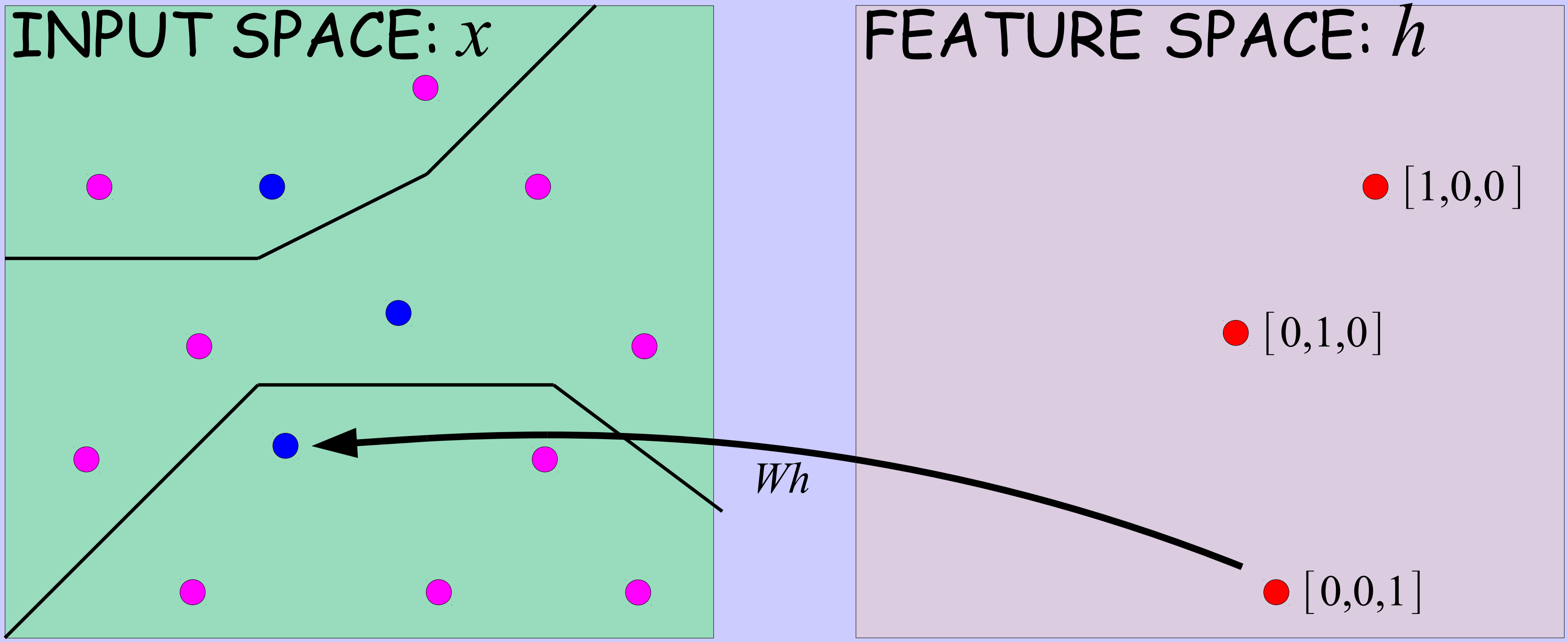


Two Approaches to Unsupervised Learning

1st strategy: constrain latent representation &
optimize score only at training samples

e.g., K-Means: score = reconstruction error: $\|x - Wh\|^2$
constraint = h 1-of-N: [0 0 0 1 0 0 … 0]

DECODING

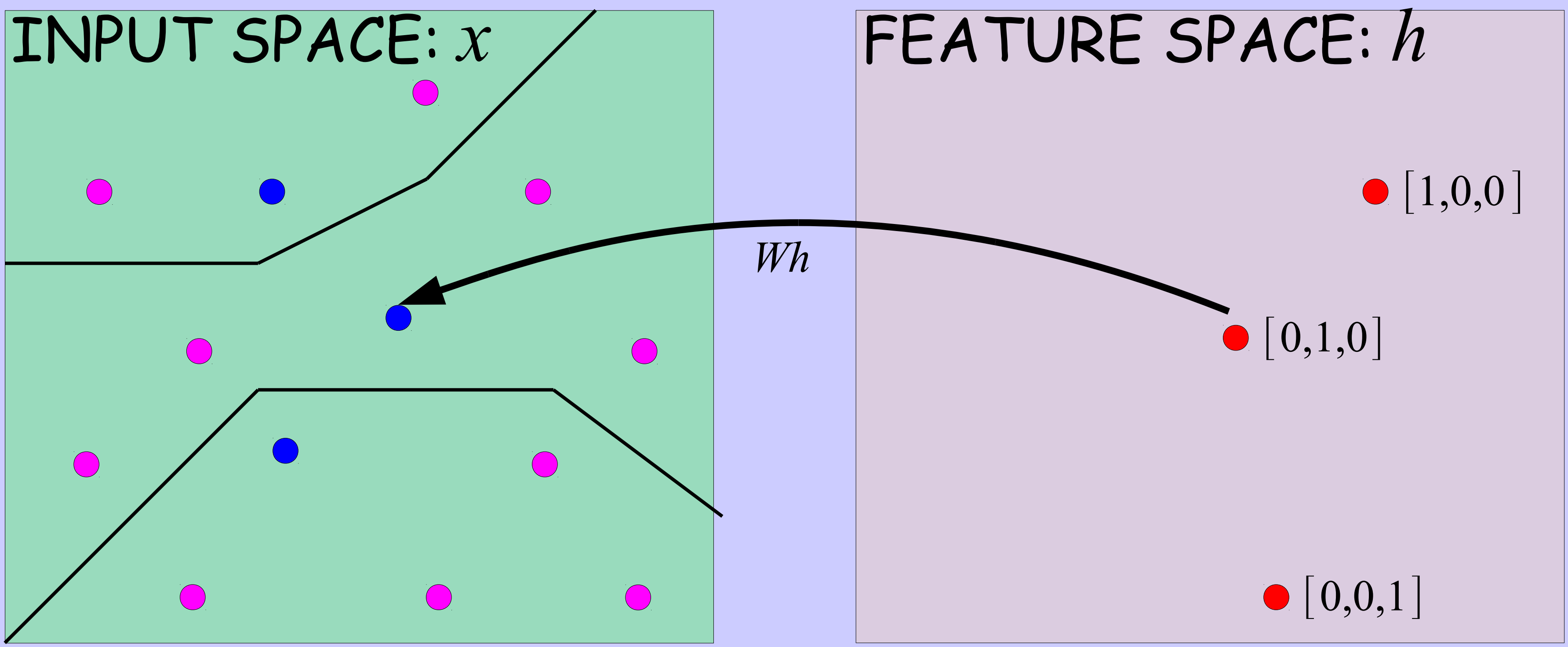


Two Approaches to Unsupervised Learning

1st strategy: constrain latent representation &
optimize score only at training samples

e.g., K-Means: score = reconstruction error: $||x - Wh||^2$
constraint = h 1-of-N: [0 0 0 1 0 0 … 0]

DECODING

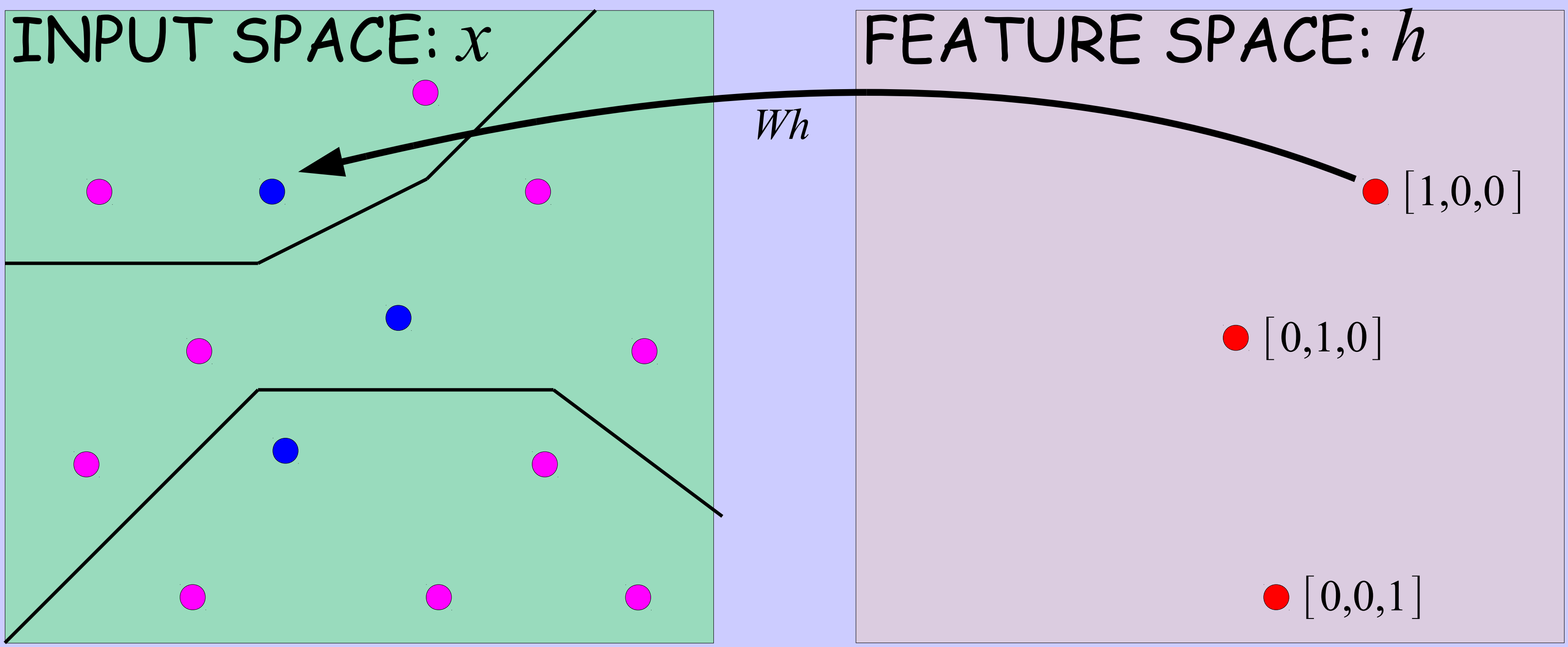


Two Approaches to Unsupervised Learning

1st strategy: constrain latent representation &
optimize score only at training samples

e.g., K-Means: score = reconstruction error: $\|x - Wh\|^2$
constraint = h 1-of-N: [0 0 0 1 0 0 … 0]

DECODING



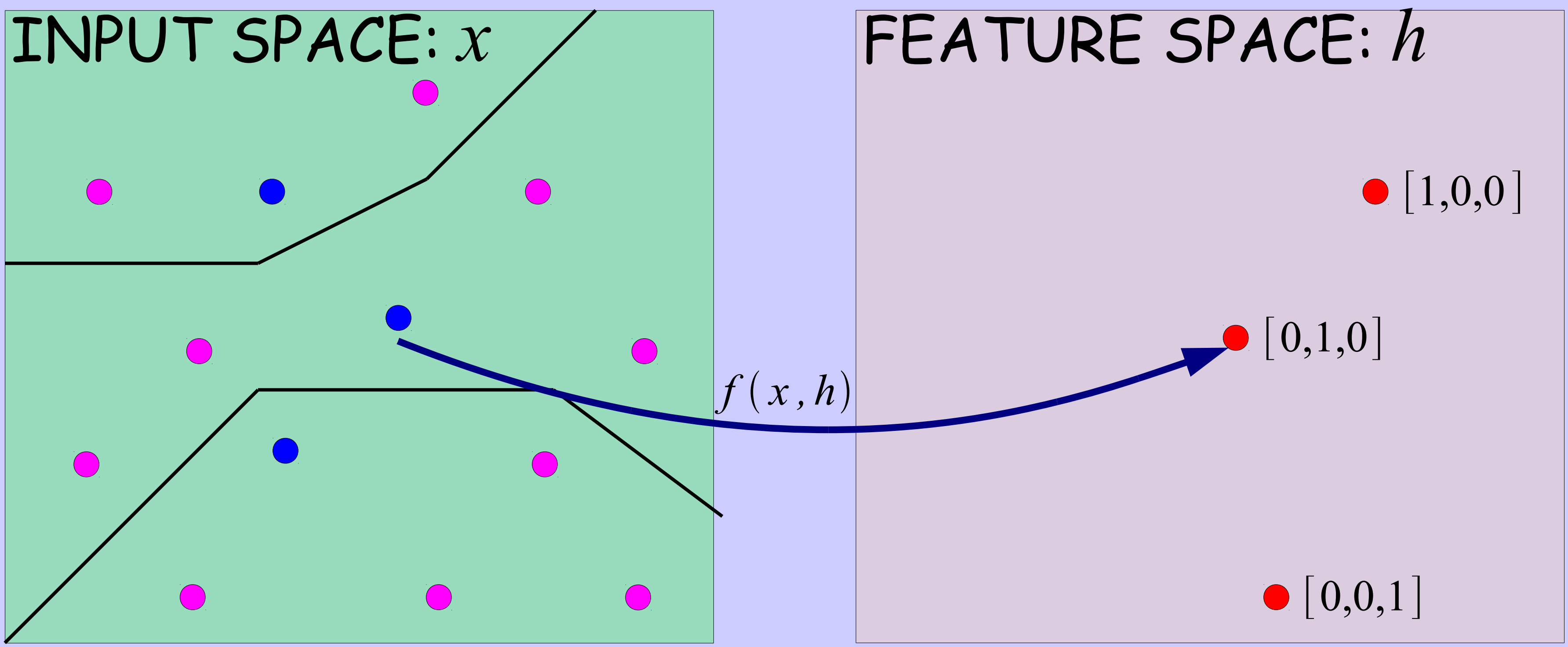
Two Approaches to Unsupervised Learning

1st strategy: constrain latent representation & optimize score only at training samples

e.g., K-Means: score = reconstruction error: $||x - Wh||^2$

constraint = h 1-of-N: [0 0 0 1 0 0 … 0]

ENCODING

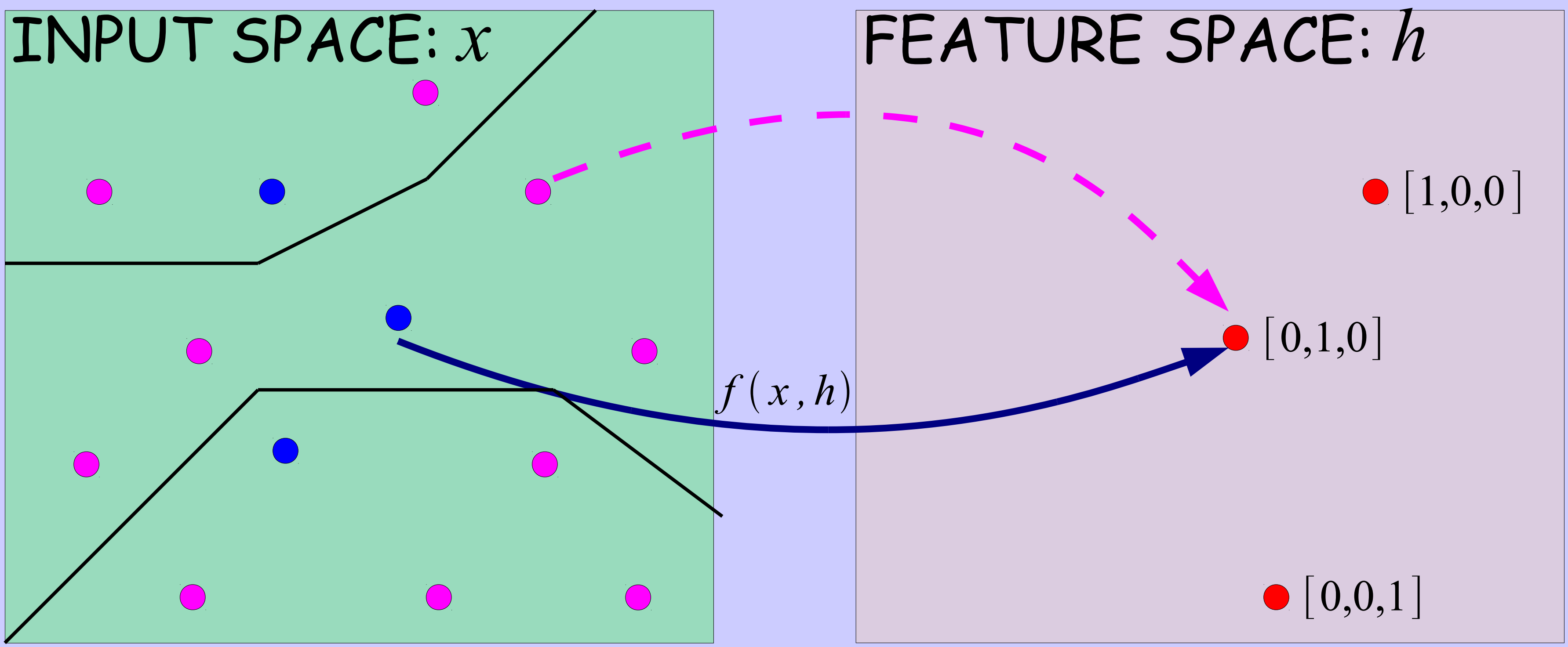


Two Approaches to Unsupervised Learning

1st strategy: constrain latent representation &
optimize score only at training samples

e.g., K-Means: score = reconstruction error: $\|x - Wh\|^2$
constraint = h 1-of-N: [0 0 0 1 0 0 ... 0]

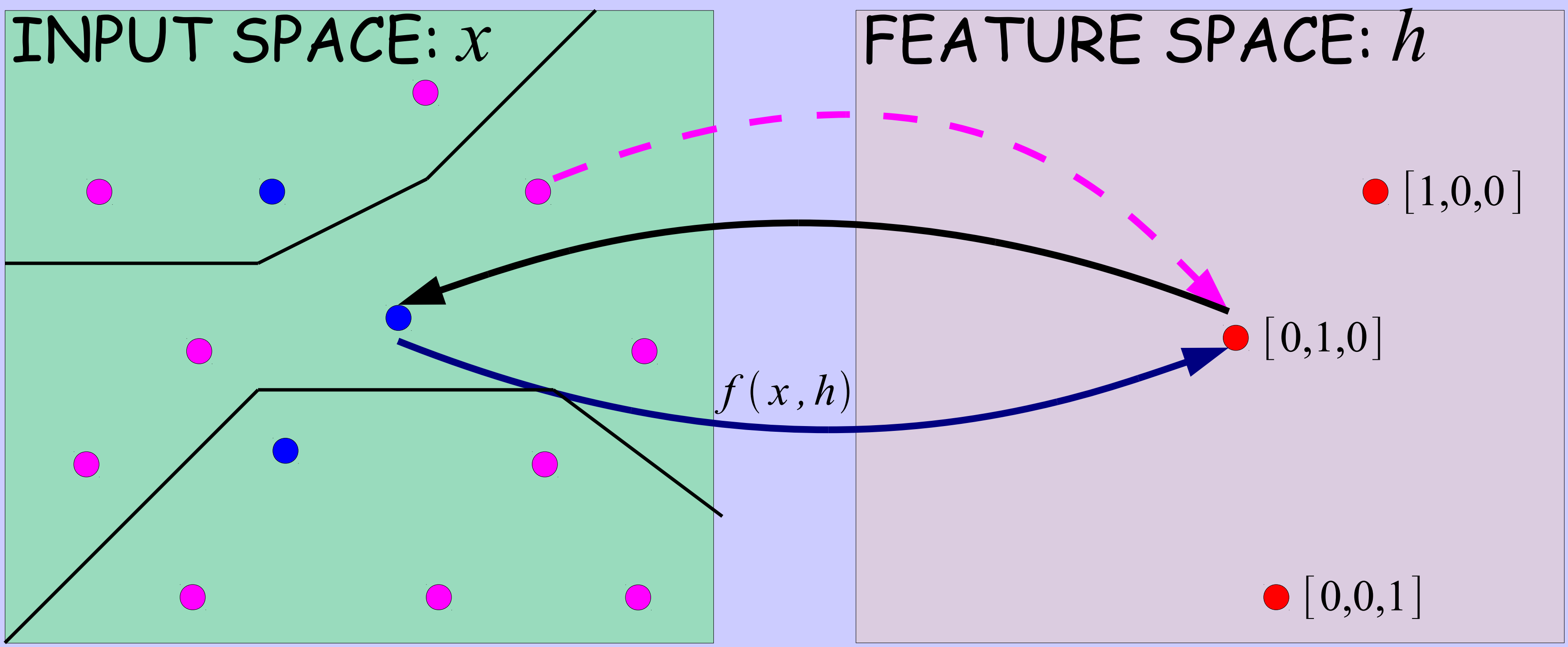
ENCODING



Two Approaches to Unsupervised Learning

1st strategy: constrain latent representation &
optimize score only at training samples

e.g., K-Means: score = reconstruction error: $\|x - Wh\|^2$
constraint = h 1-of-N: [0 0 0 1 0 0 … 0]



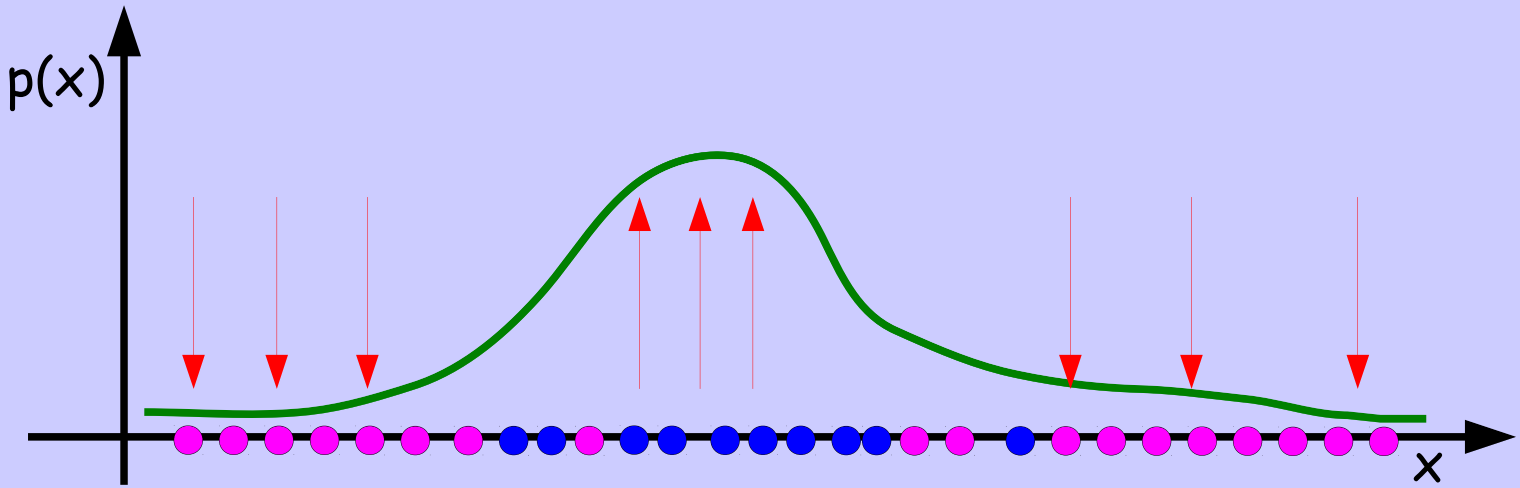
Two Approaches to Unsupervised Learning

1st strategy: constrain latent representation &
optimize score only at training samples
- K-Means
- sparse coding
- use lower dimensional representations

Two Approaches to Unsupervised Learning

1st strategy: constrain latent representation &
optimize score only at training samples
- K-Means
- sparse coding
- use lower dimensional representations

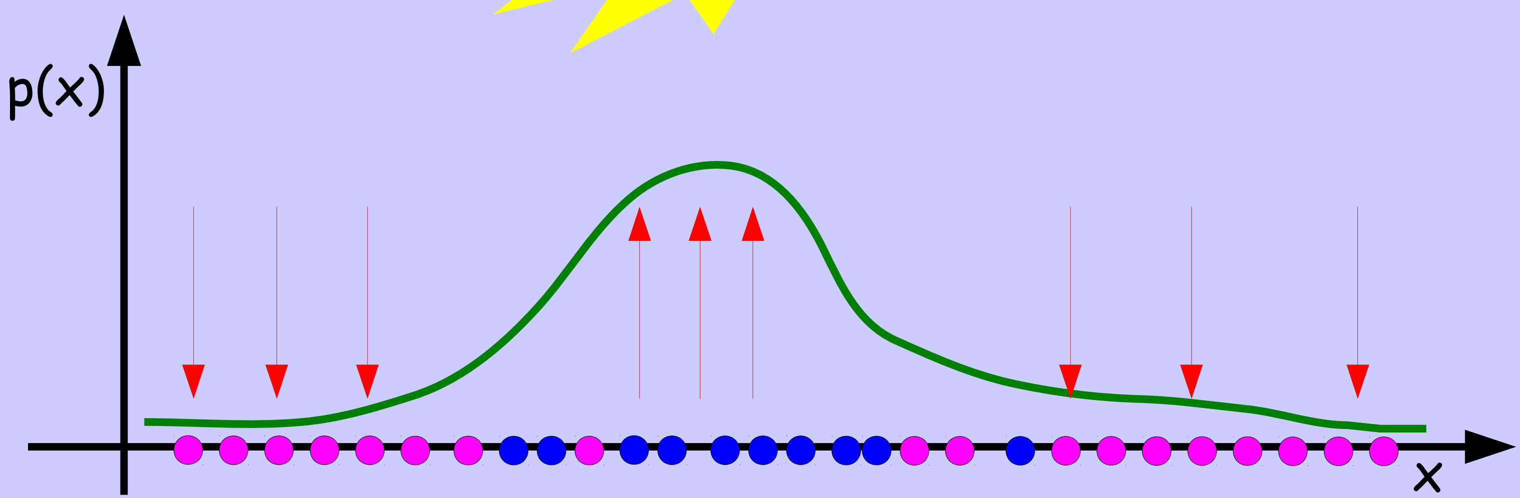
2nd strategy: optimize score for training samples while normalizing the score over the whole space (maximum likelihood)



Two Approaches to Unsupervised Learning

1st strategy: constrain latent representation &
optimize score only at training samples
- K-Means
- sparse coding
- use lower dimensional representations

2nd strategy: optimize score for samples while normalizing
the score over space (maximum likelihood)

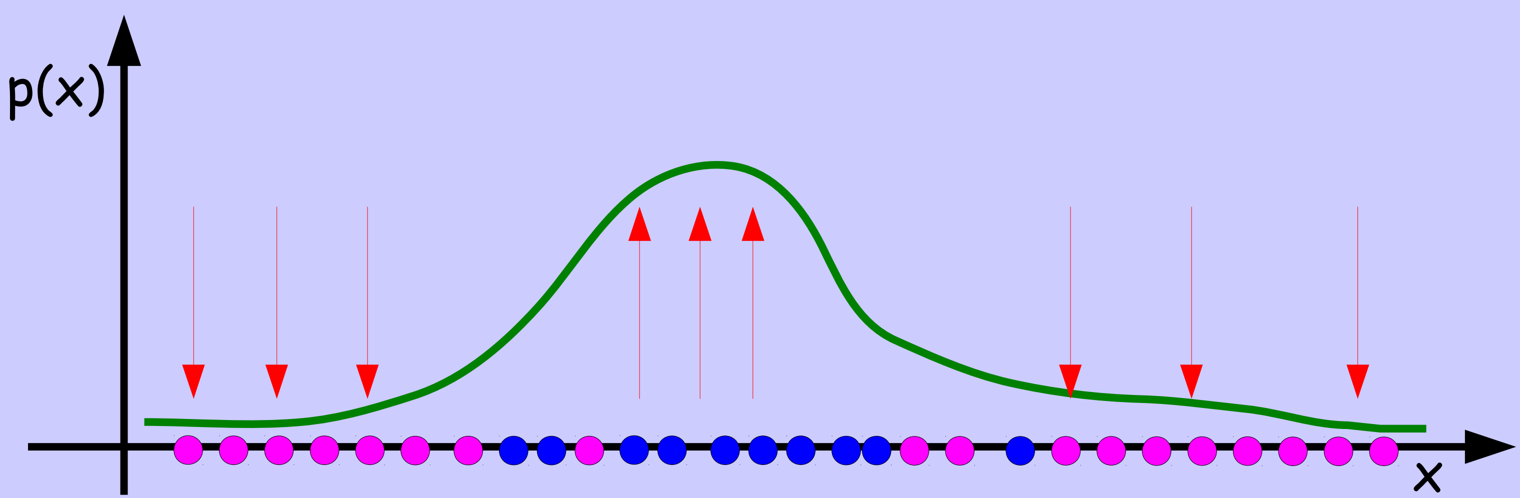


Two Approaches to Unsupervised Learning

1st strategy: constrain latent represen…
optimize score only a… …amples
- K-Means
- sparse cod…
- use lowe… …onal representations

TOMORROW

2nd strategy: opti…ize score for training samples while normalizing
the score over the whole space (maximum likelihood)



Outline

- mathematical formulation of the model
- training
- generation of natural images
- recognition of facial expression under occlusion
- learning acoustic features for spech recognition
- conclusion

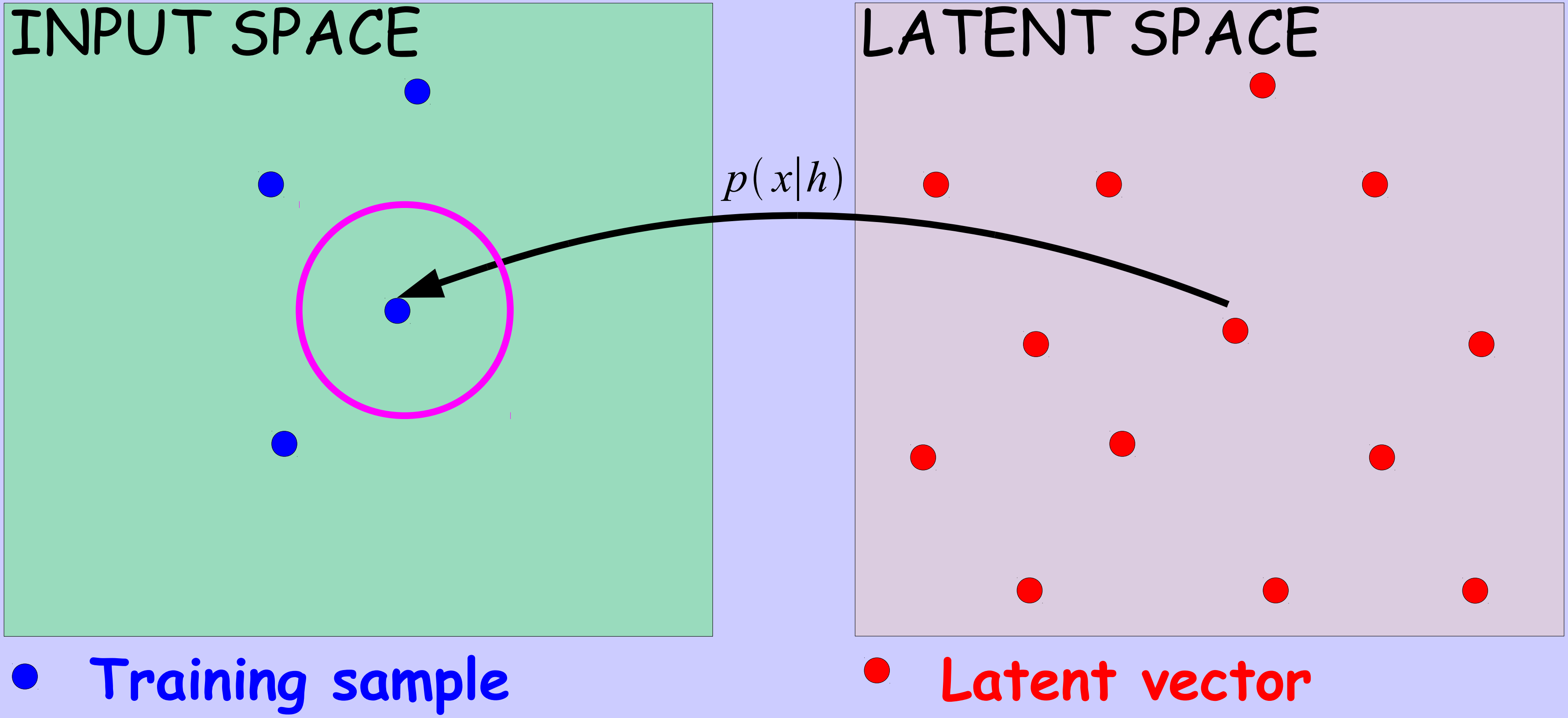
Outline

- mathematical formulation of the model
- training
- generation of natural images
- recognition of facial expression under occlusion
- learning acoustic features for spech recognition
- conclusion

Conditional Distribution Over Input

$$p(x|h) = N(mean(h), D)$$

- examples: PPCA, Factor Analysis, ICA, Gaussian RBM



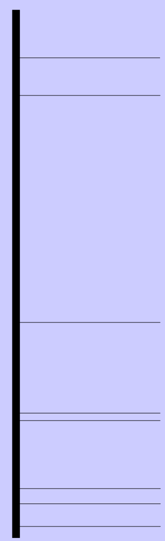
Conditional Distribution Over Input

$$p(x|h) = N(mean(h), D)$$

- examples: PPCA, Factor Analysis, ICA, Gaussian RBM

input image

$p(h|x)$



latent variables

$p(x|h)$

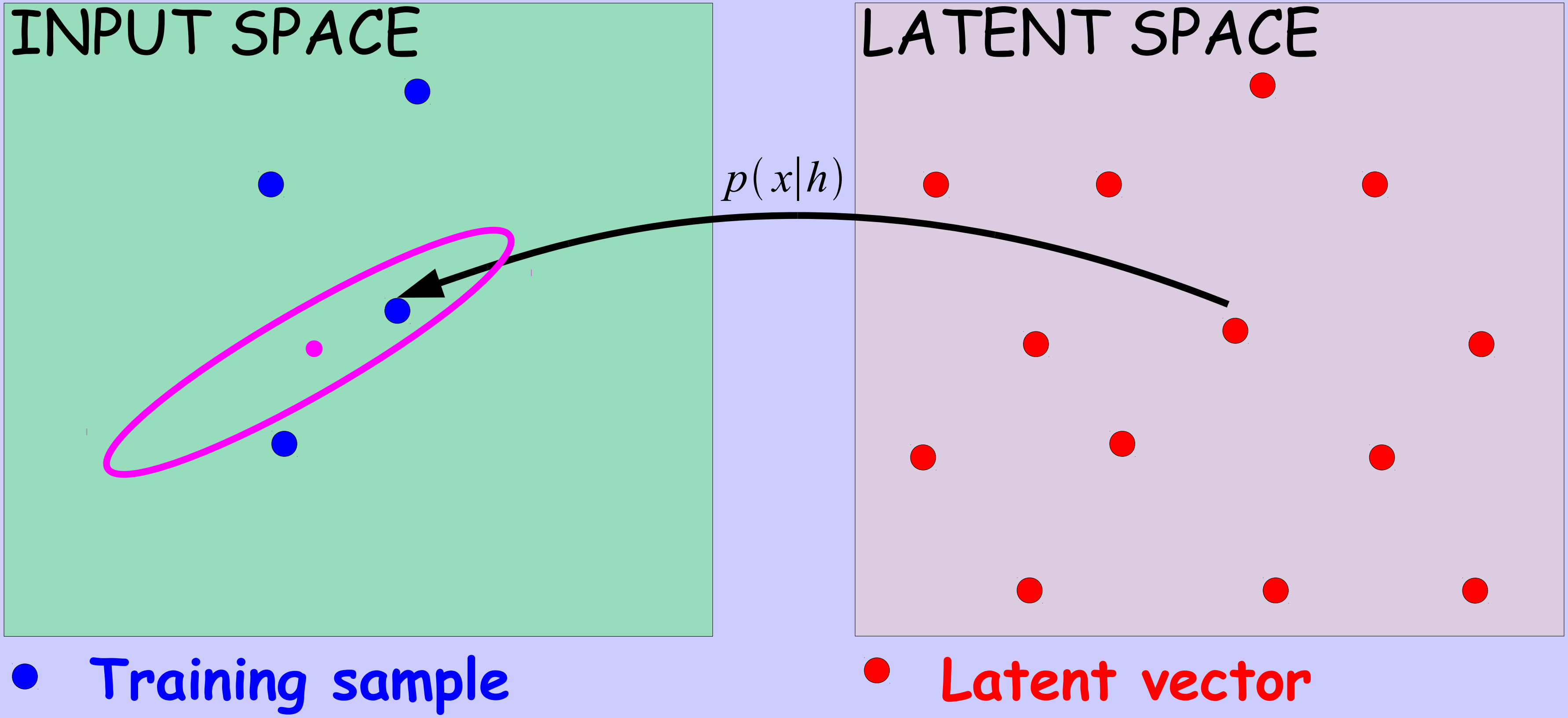
generated image

model does not represent well dependecies, only mean intensity

Conditional Distribution Over Input

$$p(x|h) = N(0, Covariance(h))$$

- examples: PoT, cRBM

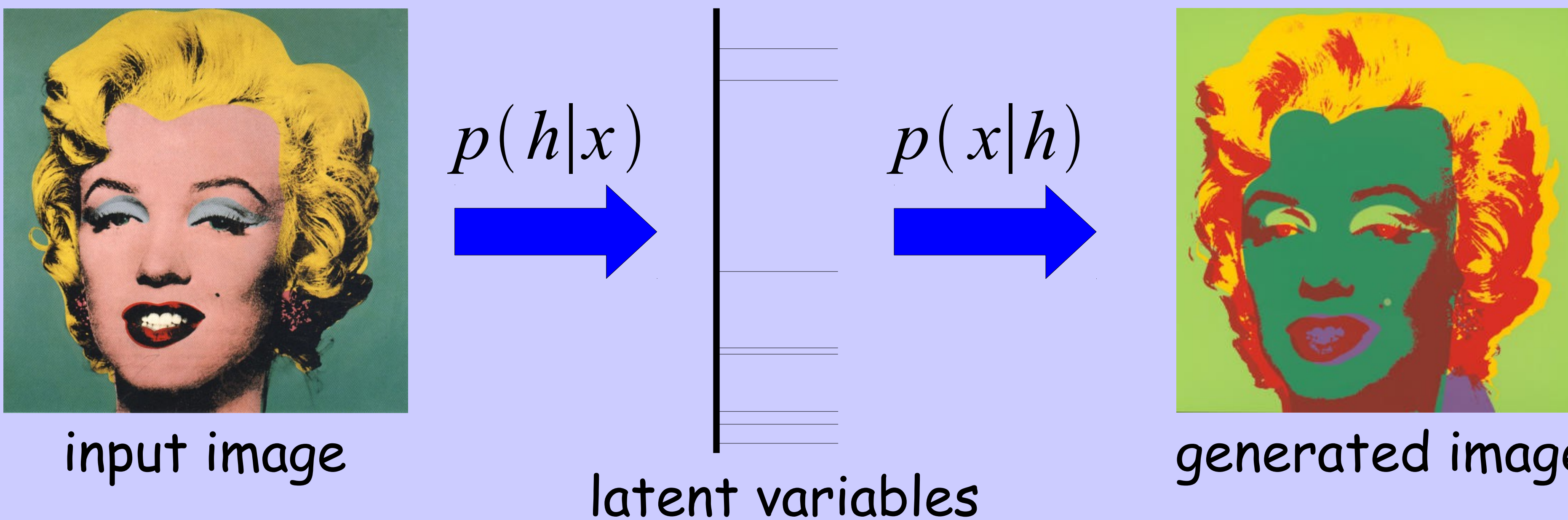


Welling et al. NIPS 2003, Ranzato et al. AISTATS 10

Conditional Distribution Over Input

$$p(x|h) = N(0, Covariance(h))$$

- examples: PoT, cRBM



model does not represent well mean intensity, only dependencies

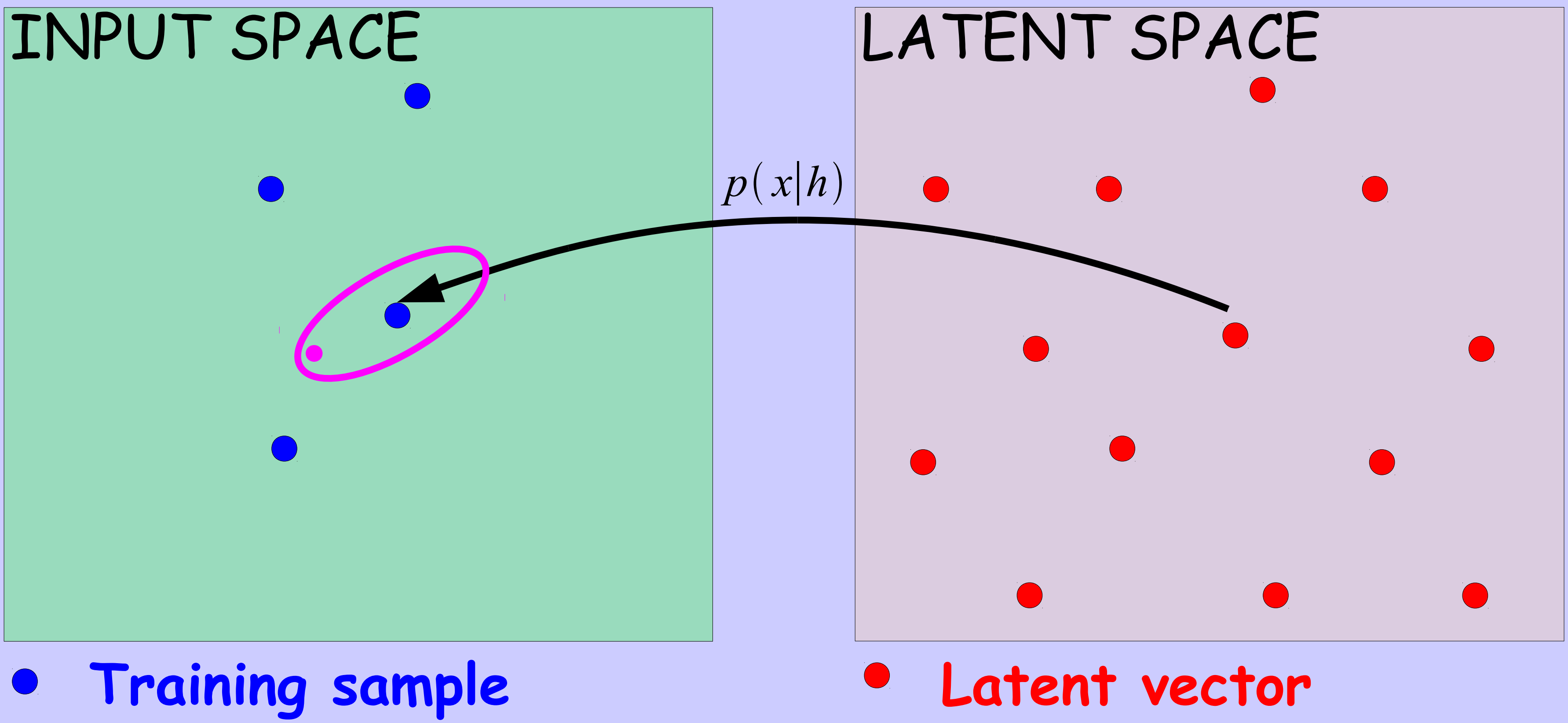
Welling et al. NIPS 2003, Ranzato et al. AISTATS 10

Andy Warhol 1960

Conditional Distribution Over Input

$$p(x|h) = N(mean(h), Covariance(h))$$

- this is what we propose: mcRBM, mPoT

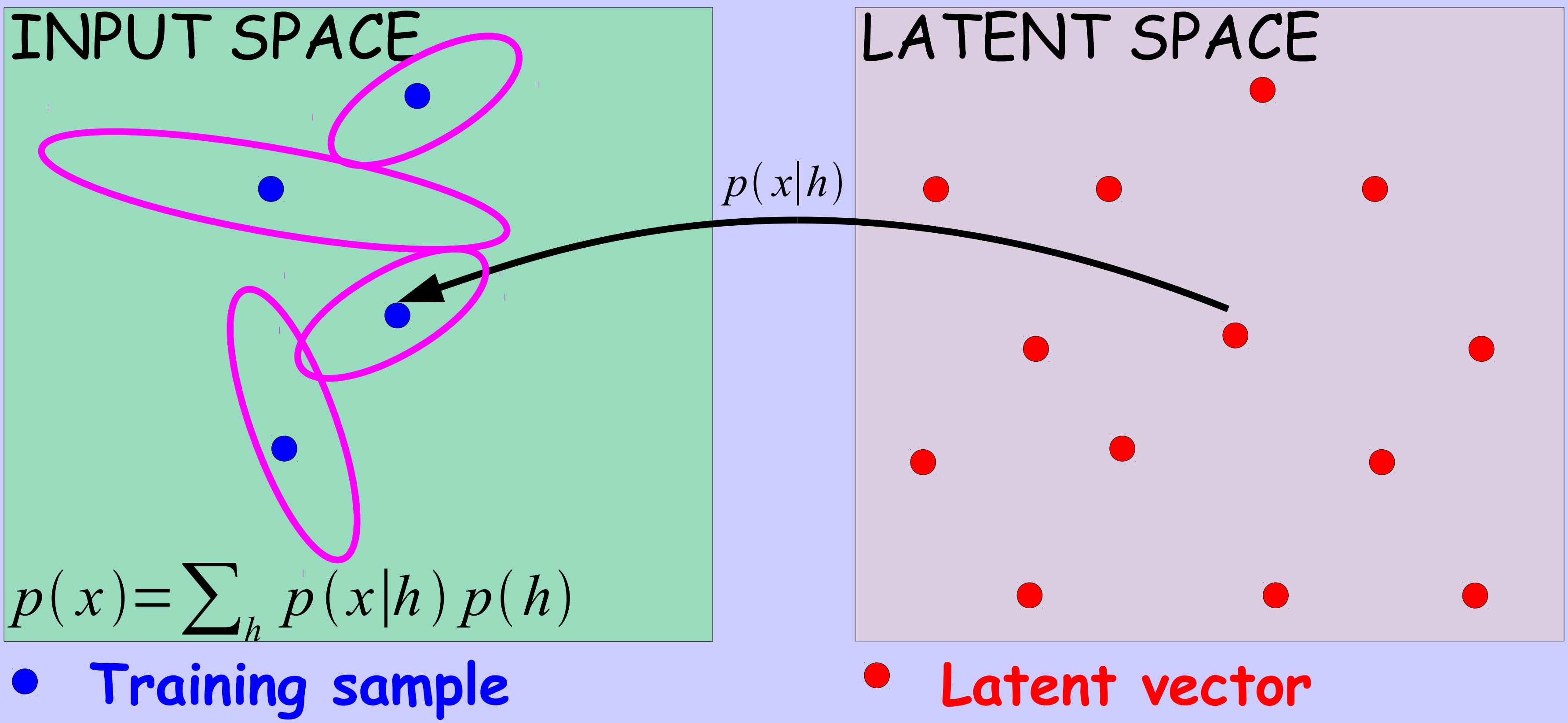


Ranzato et al. CVPR 10, Ranzato et al. NIPS 2010, Ranzato et al. CVPR 11

Conditional Distribution Over Input

$$p(x|h) = N(mean(h), Covariance(h))$$

- this is what we propose: mcRBM, mPoT



Ranzato et al. CVPR 10, Ranzato et al. NIPS 2010, Ranzato et al. CVPR 11

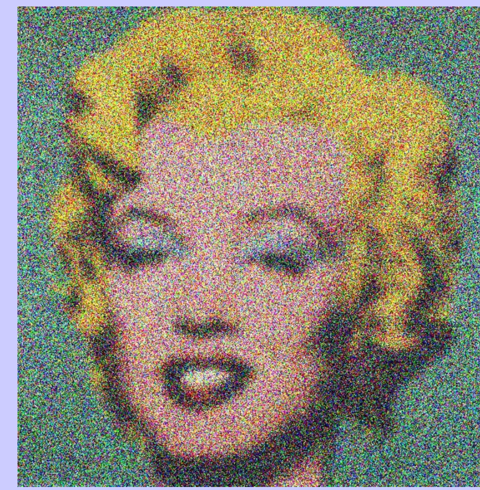
PoT

$N(0,\Sigma)$

Our model

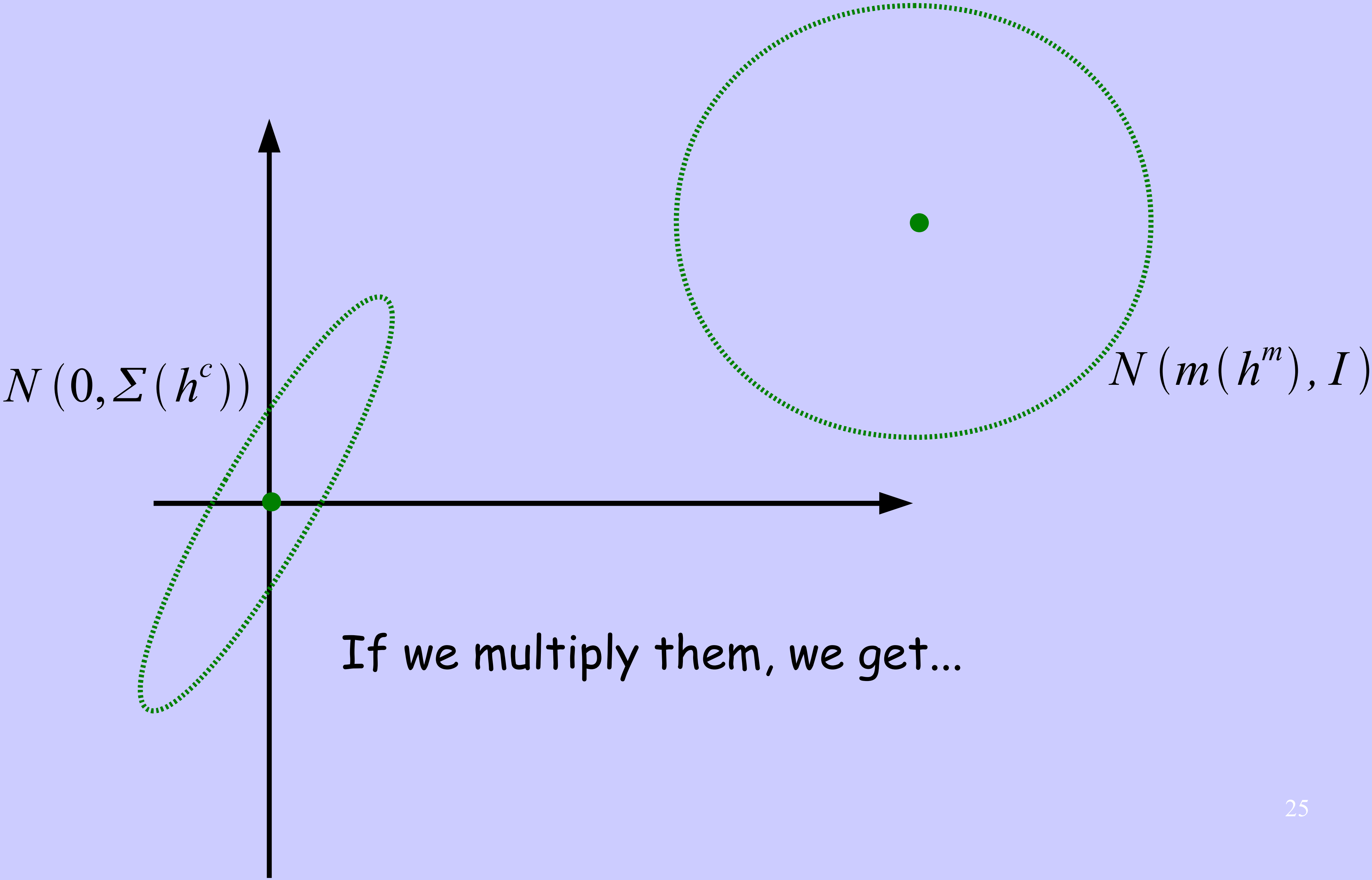
$N(m,\Sigma)$

PPCA



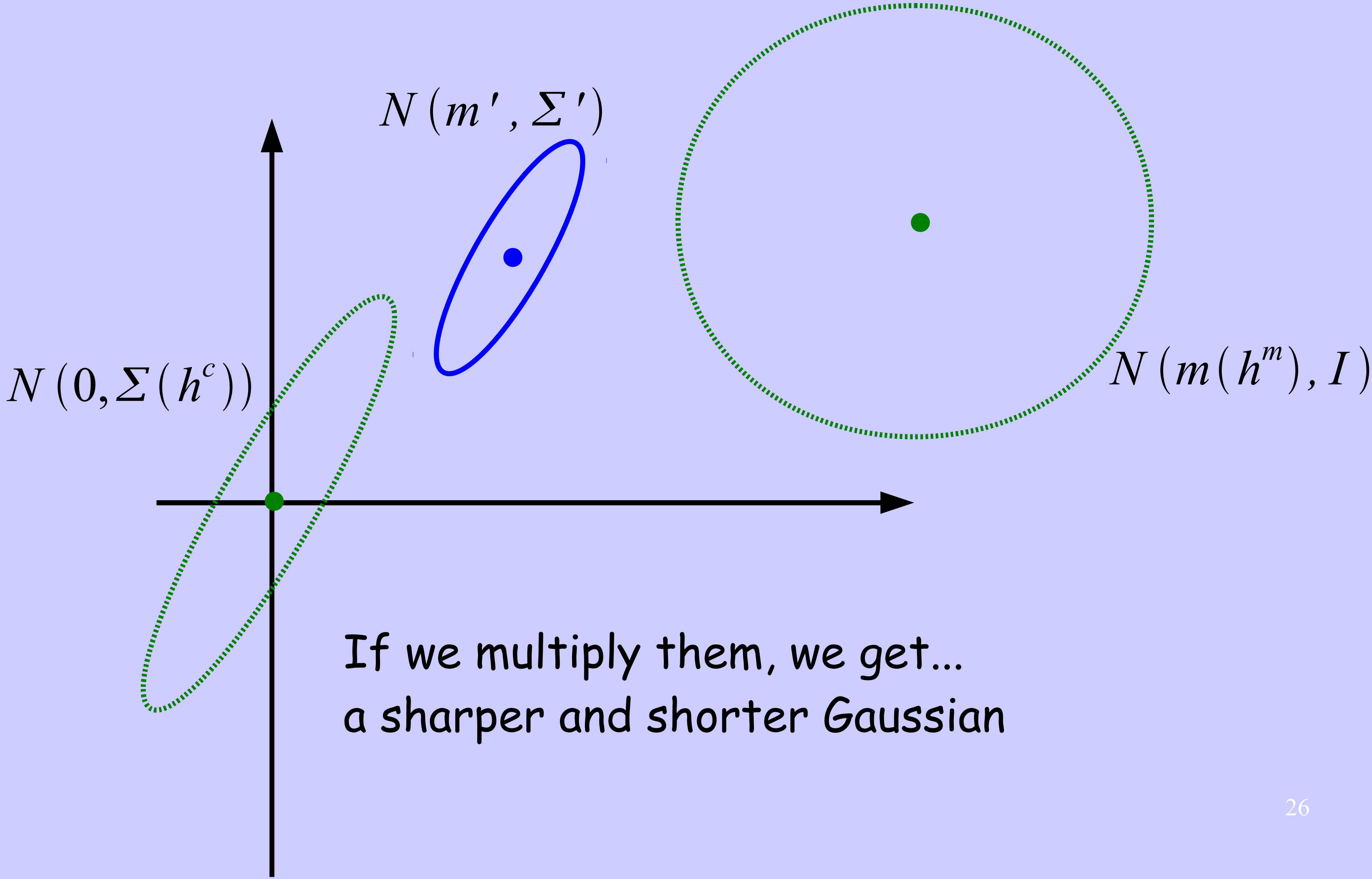
$N(m,I)$

Geometric interpretation of conditional over x

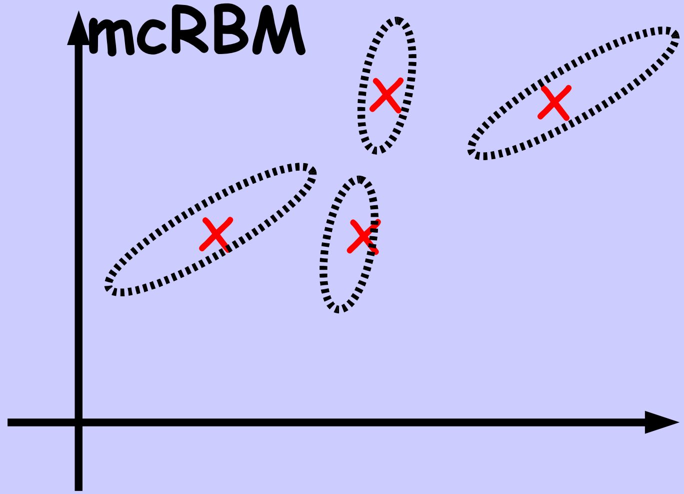


If we multiply them, we get...

Geometric interpretation of conditional over x



If we multiply them, we get...
a sharper and shorter Gaussian



- two sets of latent variables to modulate mean and covariance of the conditional distribution over the input

- energy-based model

$$p(x, h^m, h^c) \propto \exp(-E(x, h^m, h^c))$$

$$x \in \mathbb{R}^D$$
$$h^c \in \{0,1\}^M$$
$$h^m \in \{0,1\}^N$$

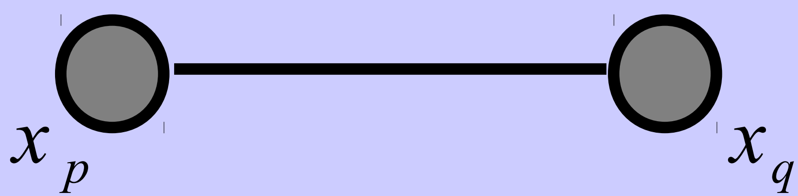
Ranzato Hinton CVPR 10

Covariance part of the energy function:

$$E(x, h^c, h^m) = \frac{1}{2} x' \Sigma^{-1} x$$

$x \in \mathbb{R}^D$

$\Sigma^{-1} \in \mathbb{R}^{D \times D}$



pair-wise MRF

Ranzato Hinton CVPR 10

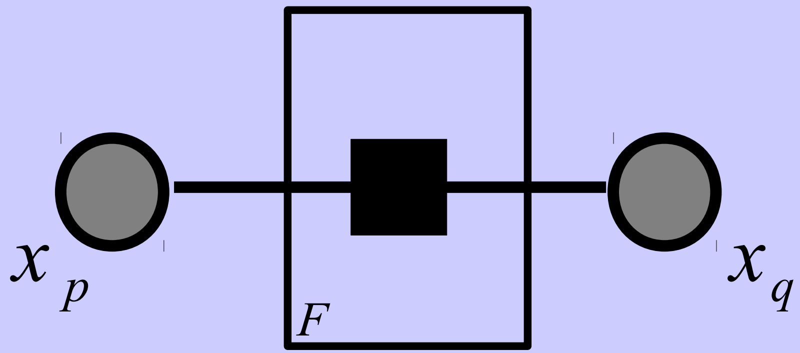
Covariance part of the energy function:

$$E(x, h^c, h^m) = \frac{1}{2} x' C C' x$$

$x \in \mathbb{R}^D$

factorization

$C \in \mathbb{R}^{D \times F}$



pair-wise MRF

Ranzato Hinton CVPR 10

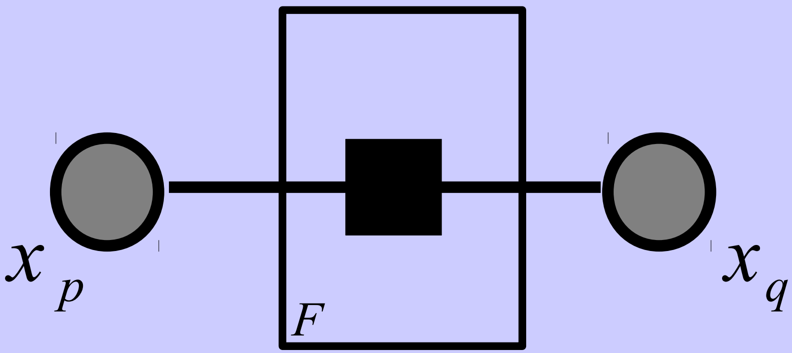
Covariance part of the energy function:

$$E(x, h^c, h^m) = \frac{1}{2} x' C C' x$$

$x \in \mathbb{R}^D$

factorization

$C \in \mathbb{R}^{D \times F}$



pair-wise MRF

$$E(x, h^c, h^m) = \frac{1}{2} x' C C' x = \alpha_{11} x_1^2 + \alpha_{12} x_1 x_2 + \dots$$

Ranzato Hinton CVPR 10

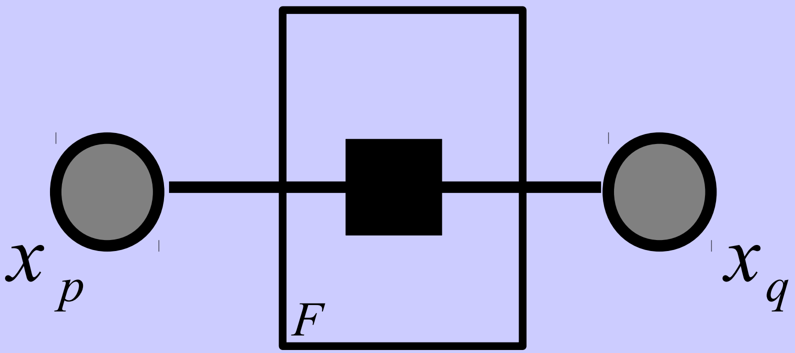
Covariance part of the energy function:

$$E(x, h^c, h^m) = \frac{1}{2} x' C C' x$$

$x \in \mathbb{R}^D$

$C \in \mathbb{R}^{D \times F}$

factorization



pair-wise MRF

$$E(x, h^c, h^m) = \frac{1}{2} x' C C' x = \frac{1}{2} \sum_{i=1}^{F} (C_i' x)^2$$

Ranzato Hinton CVPR 10

Covariance part of the energy function:

$$E(x, h^c, h^m) = \frac{1}{2} x' C [diag(h^c)] C' x$$

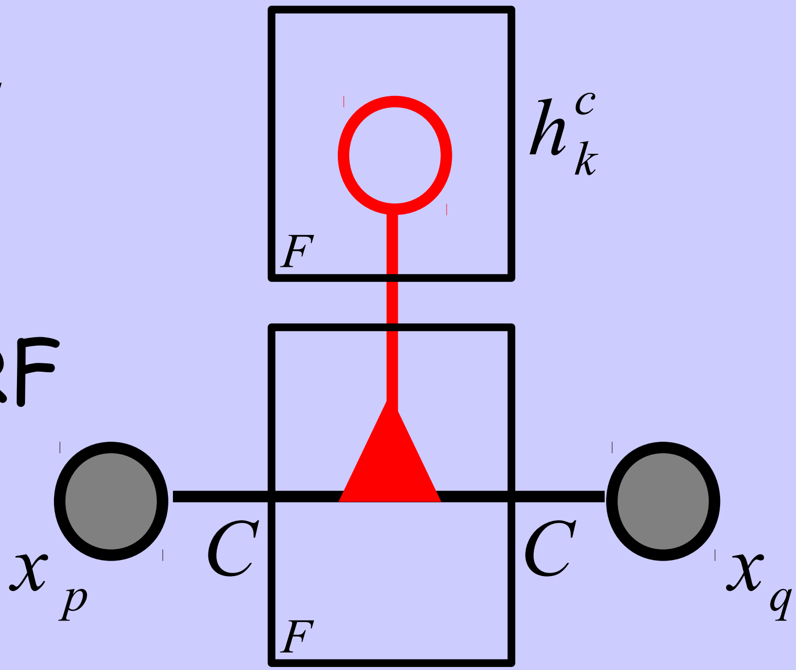
$x \in \mathbb{R}^D$

$C \in \mathbb{R}^{D \times F}$

$h^c \in \{0,1\}^F$

factorization + hiddens

gated MRF



Ranzato Hinton CVPR 10

Covariance part of the energy function:

$$E(x, h^c, h^m) = \frac{1}{2} x' C [diag(h^c)] C' x$$

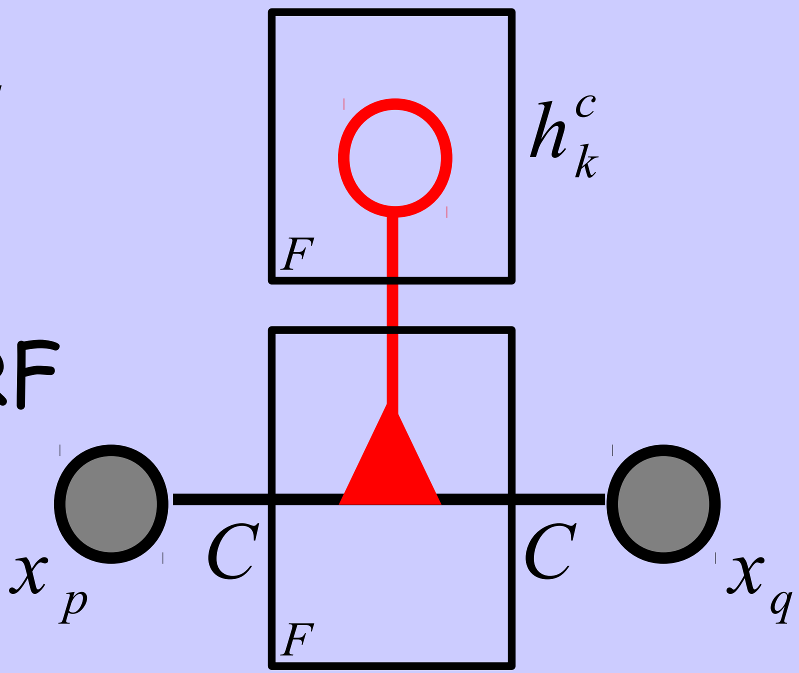
$x \in \mathbb{R}^D$

$C \in \mathbb{R}^{D \times F}$

$h^c \in \{0,1\}^F$

factorization + hiddens

gated MRF



$$E(x, h^c, h^m) = \frac{1}{2} x' C [diag(h^c)] C' x = \frac{1}{2} \sum_{i=1}^{F} h_i^c (C_i' x)^2$$

Ranzato Hinton CVPR 10

Covariance part of the energy function:

$$E(x, h^c, h^m) = \frac{1}{2} x' C [diag(P h^c)] C' x$$

$x \in \mathbb{R}^D$

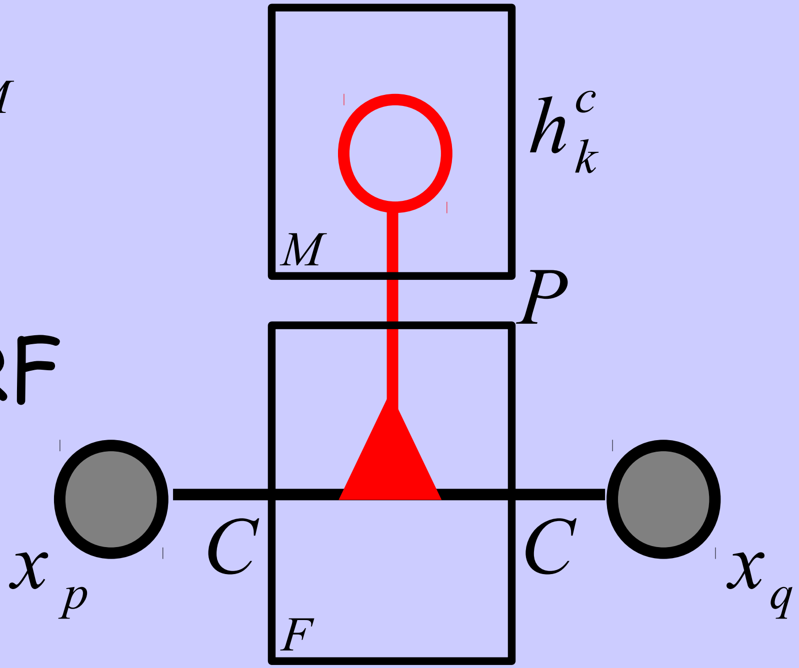
$C \in \mathbb{R}^{D \times F}$

$h^c \in \{0,1\}^M$

$P \in \mathbb{R}^{F \times M}$

factorization + hiddens

gated MRF



Ranzato Hinton CVPR 10

Covariance part of the energy function:

$$E(x, h^c, h^m) = \frac{1}{2} x' C [diag(P h^c)] C' x$$

$x \in \mathbb{R}^D$

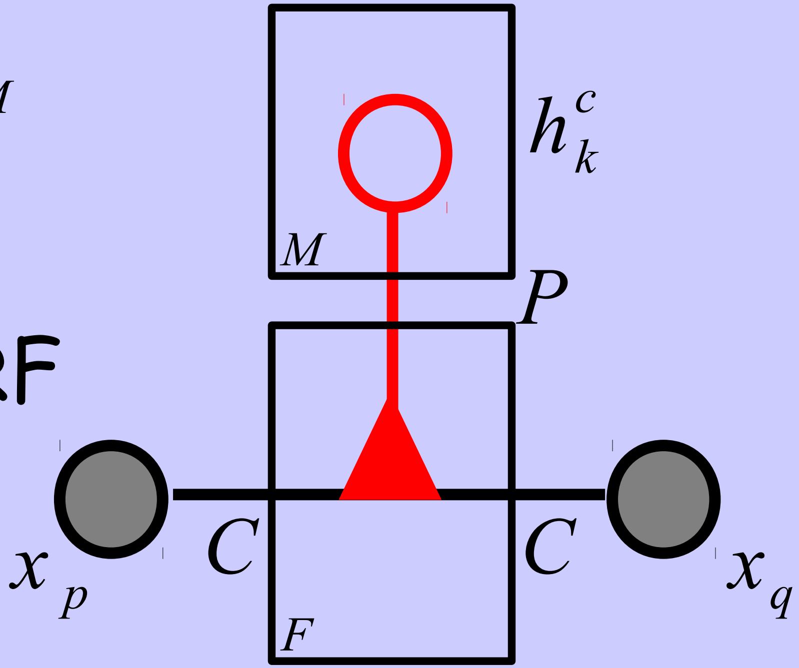
$C \in \mathbb{R}^{D \times F}$

$h^c \in \{0,1\}^M$

$P \in \mathbb{R}^{F \times M}$

factorization + hiddens

gated MRF



$$E(x, h^c, h^m) = \frac{1}{2} \sum_{k=1}^{M} \sum_{i=1}^{F} h_k^c P_{ik} (C_i' x)^2$$

Ranzato Hinton CVPR 10

Overall energy function:

$$E(x, h^c, h^m) = \frac{1}{2} x' C [diag(P h^c)] C' x + \frac{1}{2} x' x - x' W h^m$$

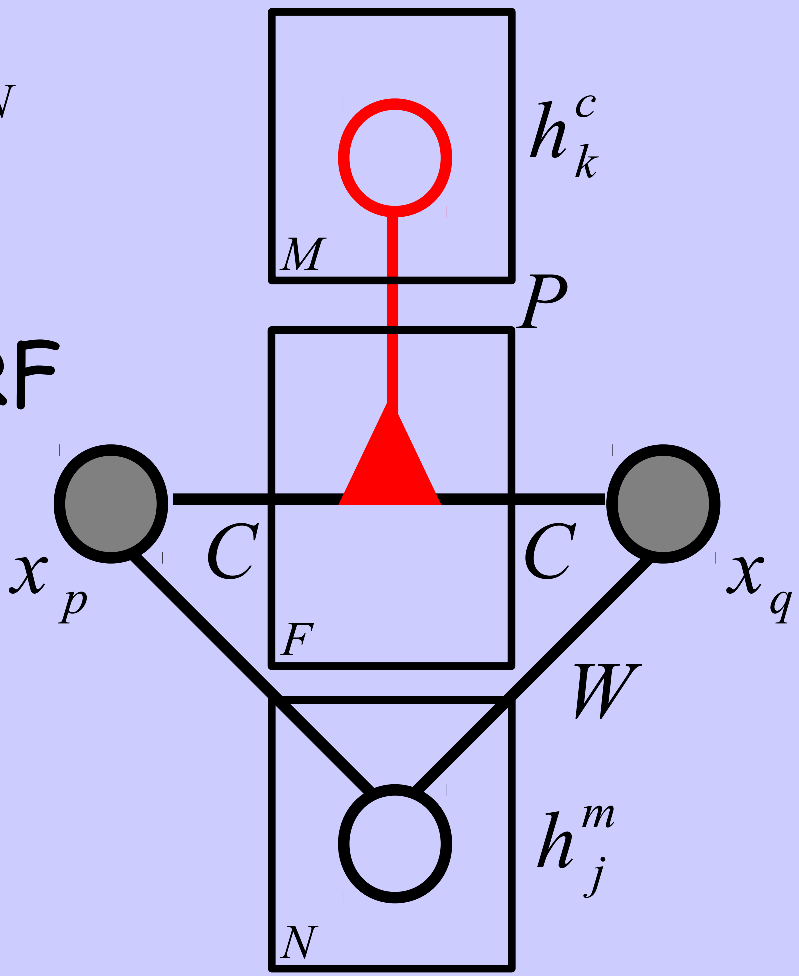
covariance part — mean part

$x \in \mathbb{R}^D$

$W \in \mathbb{R}^{D \times N}$

$h^m \in \{0,1\}^N$

gated MRF



Ranzato Hinton CVPR 10

Overall energy function:

$$E(x, h^c, h^m) = \frac{1}{2} x' C [diag(P h^c)] C' x + \frac{1}{2} x' x - x' W h^m$$

covariance part

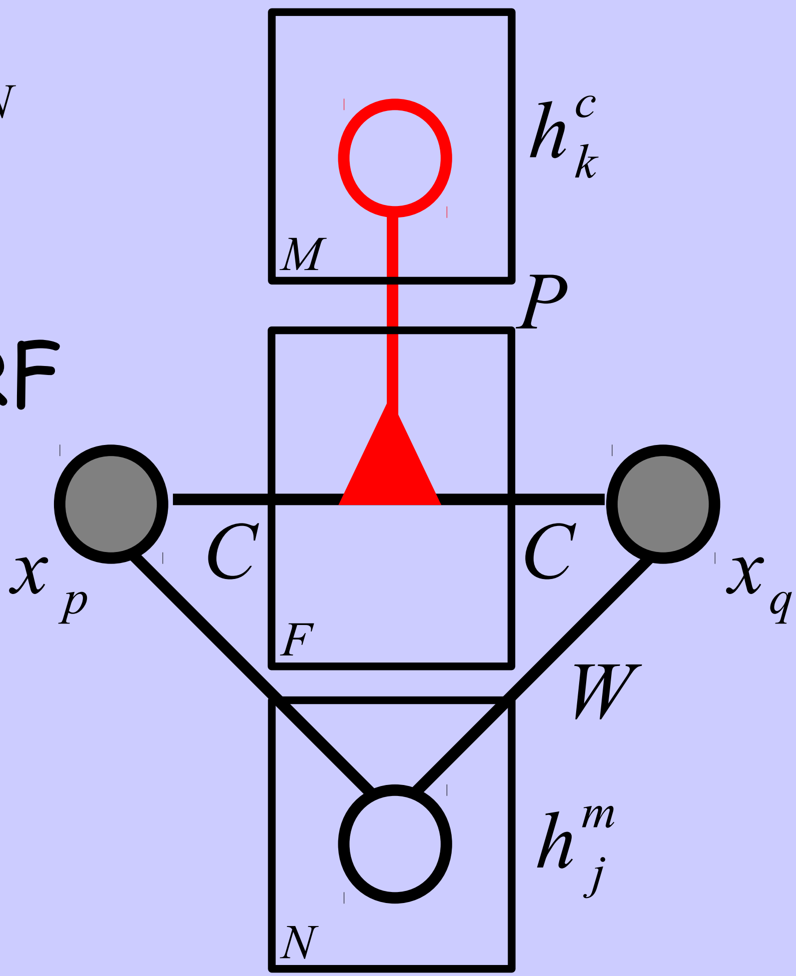
mean part

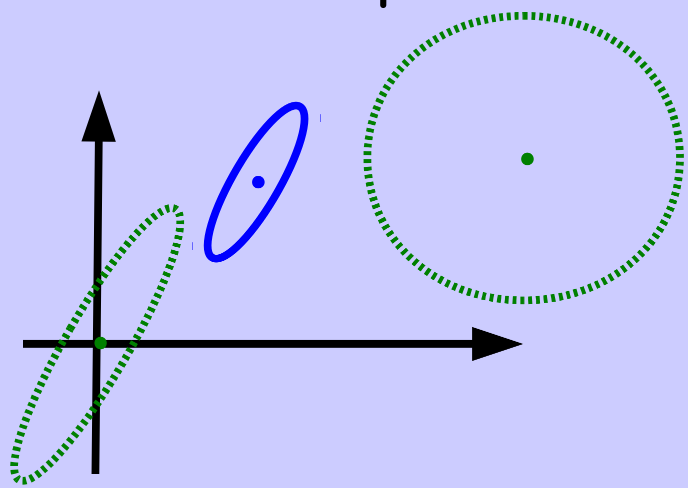
$x \in \mathbb{R}^D$

$W \in \mathbb{R}^{D \times N}$

$h^m \in \{0,1\}^N$

gated MRF





Ranzato Hinton CVPR 10

Overall energy function:

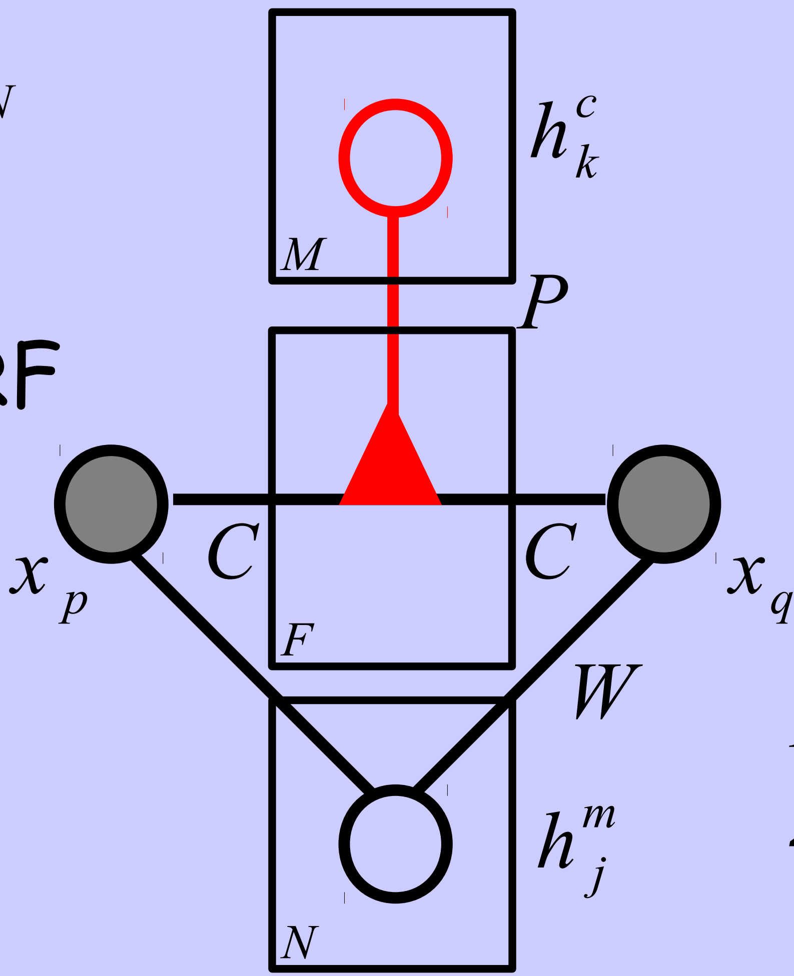
$$E(x, h^c, h^m) = \underbrace{\frac{1}{2} x' C [diag(P h^c)] C' x}_{\text{covariance part}} + \underbrace{\frac{1}{2} x' x - x' W h^m}_{\text{mean part}}$$

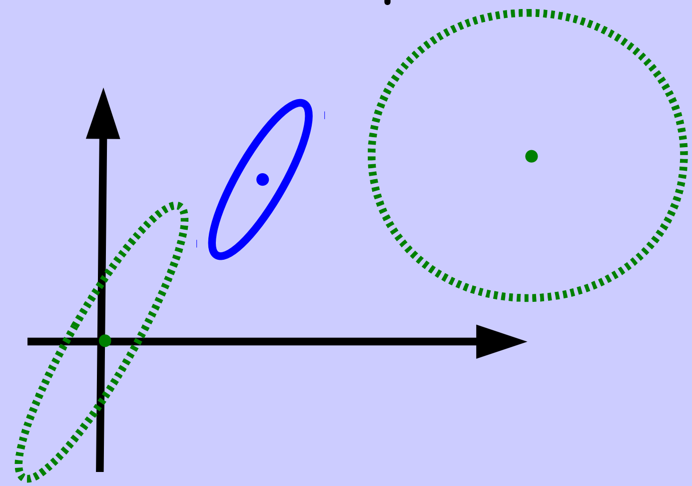
$x \in \mathbb{R}^D$

$W \in \mathbb{R}^{D \times N}$

$h^m \in \{0,1\}^N$

gated MRF





$$p(x | h^c, h^m) = N(\Sigma(W h^m), \Sigma)$$

$$\Sigma^{-1} = C \, diag[P h^c] C' + I$$

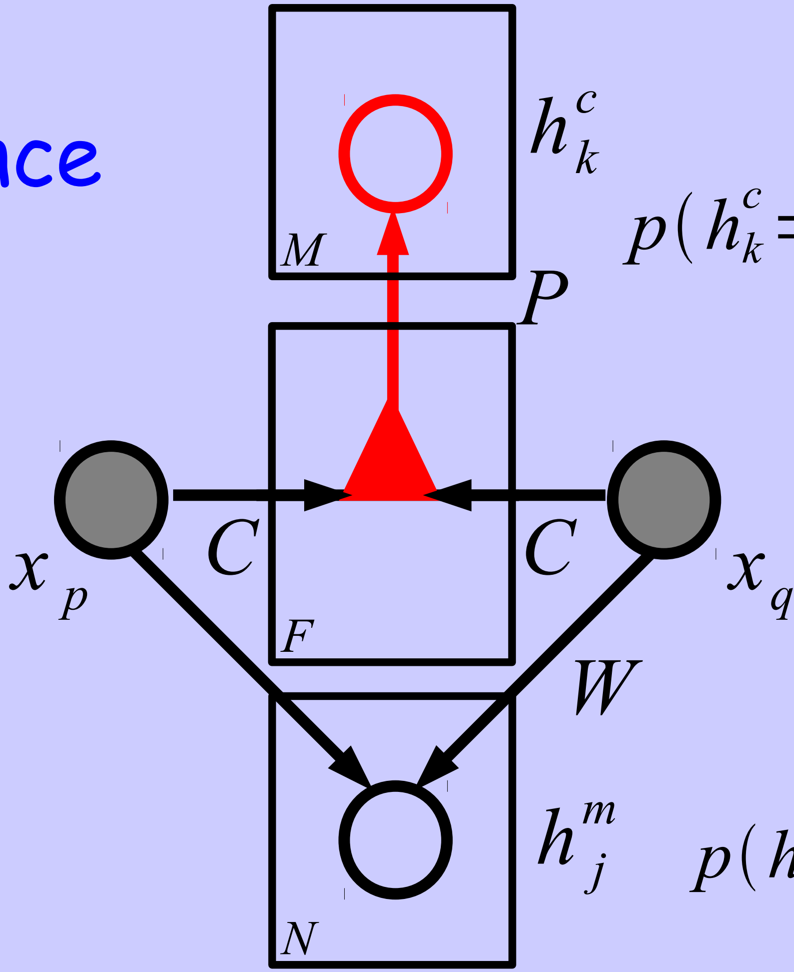
Ranzato Hinton CVPR 10

Overall energy function:

$$E(x, h^c, h^m) = \frac{1}{2} x' C [diag(P h^c)] C' x + \frac{1}{2} x' x - x' W h^m$$

covariance part mean part

inference



$$p(h^c_k = 1 | x) = \sigma(-\frac{1}{2} P_k (C' x)^2 + b_k)$$

$$p(h^m_j = 1 | x) = \sigma(W_j' x + b_j)$$

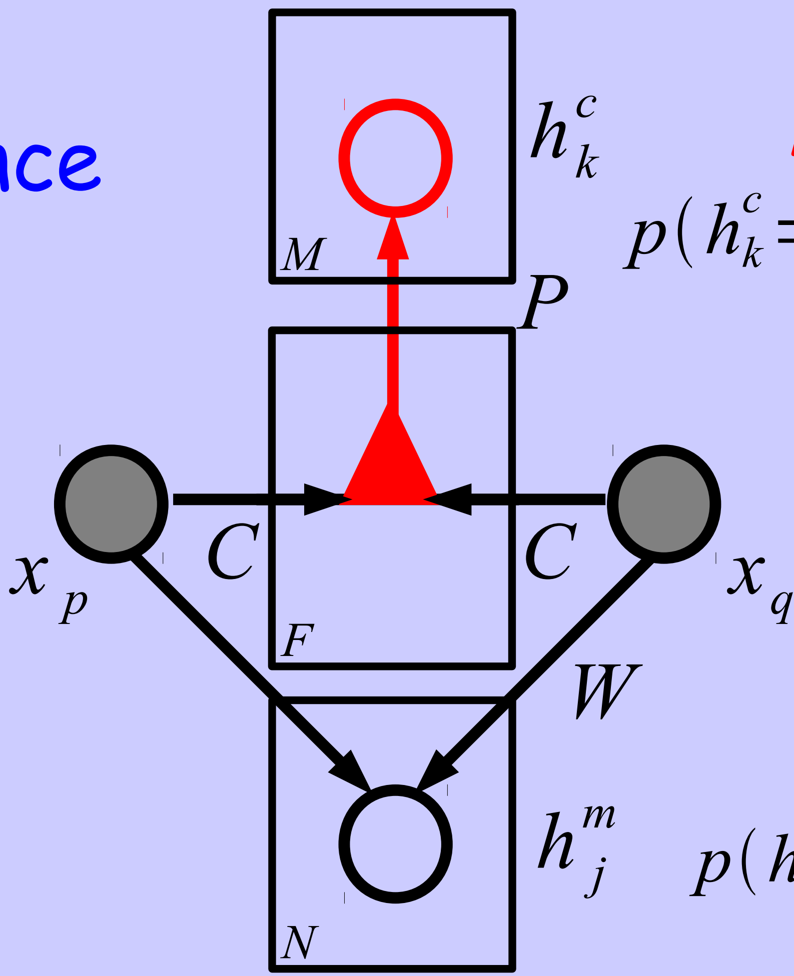
Ranzato Hinton CVPR 10

Overall energy function:

$$E(x,h^c,h^m)=\frac{1}{2}x'C[diag(Ph^c)]C'x+\frac{1}{2}x'x-x'Wh^m$$

covariance part mean part

inference



Complex-cell:

pools rectified simple cells

$$p(h_k^c=1|x)=\sigma(-\frac{1}{2}P_k(C'x)^2+b_k)$$

Simple-cell:

non-linear filtering

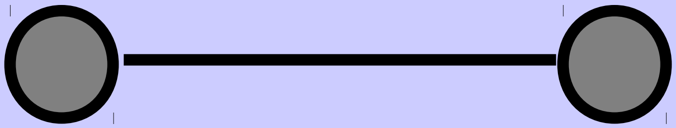
$$p(h_j^m=1|x)=\sigma(W_j{}'x+b_j)$$

Ranzato Hinton CVPR 10

Interpretation

$$E = (w'x)^2$$

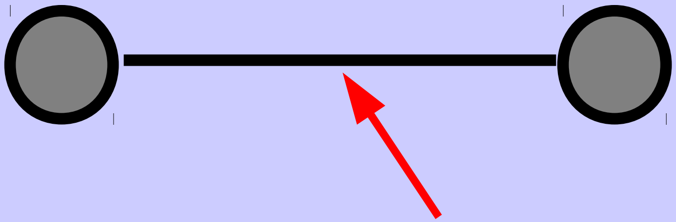
minimizing E over the training set yields the minor component: w = [-1,1] since images are usually smooth.



Interpretation

$$E = (w'x)^2$$

minimizing E over the training set yields the minor component: w = [-1,1] since images are usually smooth.



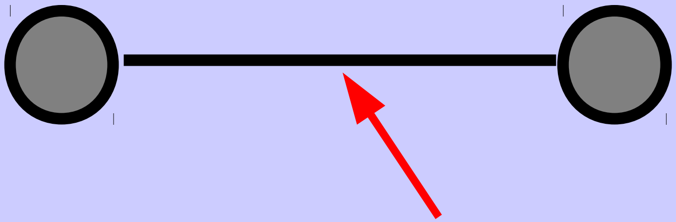
This edge shows the strong dependency (correlation) between image pixels!

Interpretation

$$E = (w'x)^2$$

This enforces a strong penalty against the violation of the constraint:

$$x_1 = x_2$$



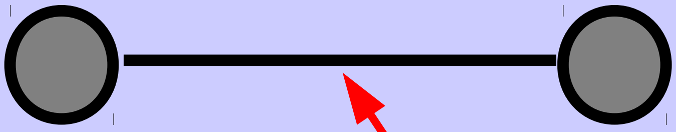
This edge shows the strong dependency (correlation) between image pixels!

Interpretation

$$E = (w'x)^2$$

How to make the penalty less strong?

How to model violations of the constraint?



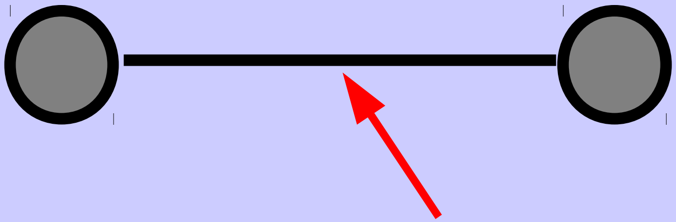
This edge shows the strong dependency (correlation) between image pixels!

Interpretation

$$E = (w'x)^2$$

How to make the penalty less strong?
How to model violations of the constraint?

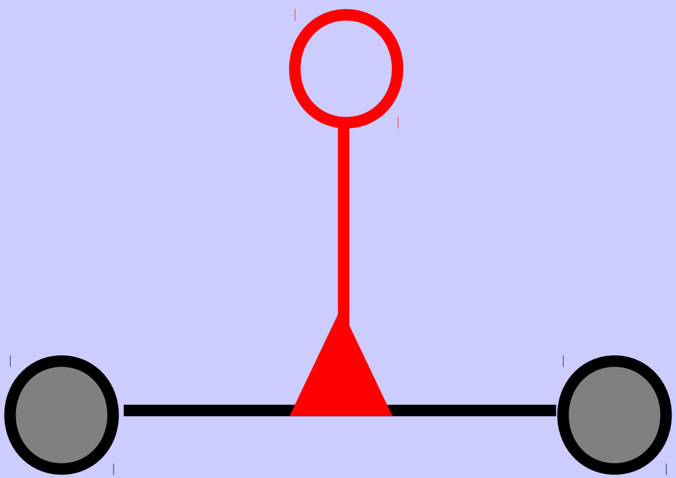
ADD LATENT VARIABLES!



This edge shows the strong dependency (correlation) between image pixels!

Interpretation

$$E = h(w'x)^2 - bh, \;\; b > 0$$



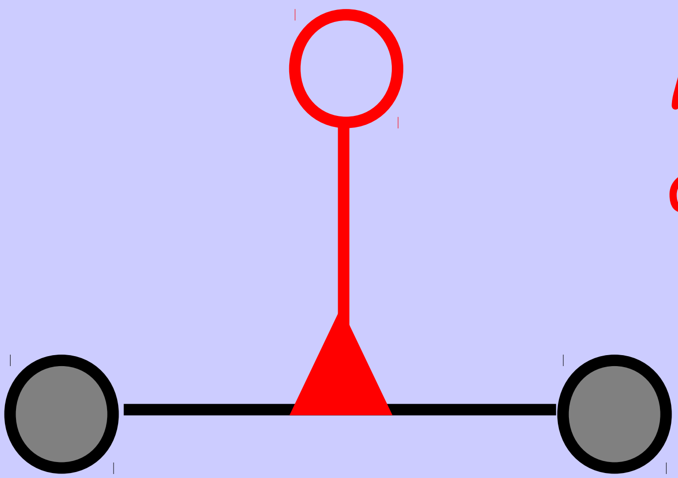
$$w'x = 0, \;\; h = 1$$
$$E = -b$$

$$w'x \gg 0, \;\; h = 0$$
$$E = 0$$

Penalty discount!

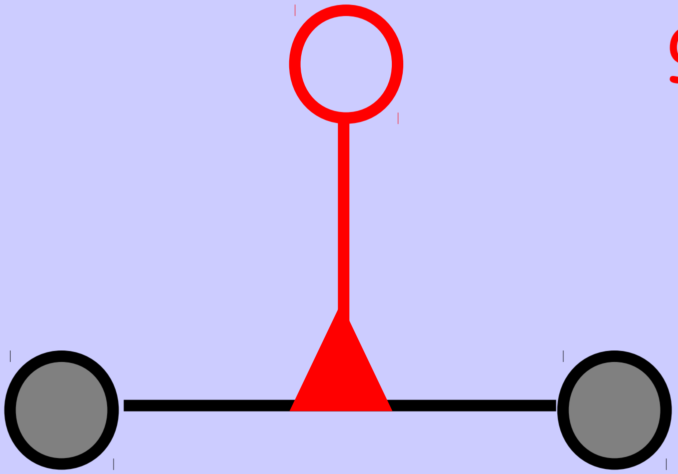
Black Rangarajan "On..line process.." IJCV 96

Interpretation



MRF with adaptive (input-dependent) affinities

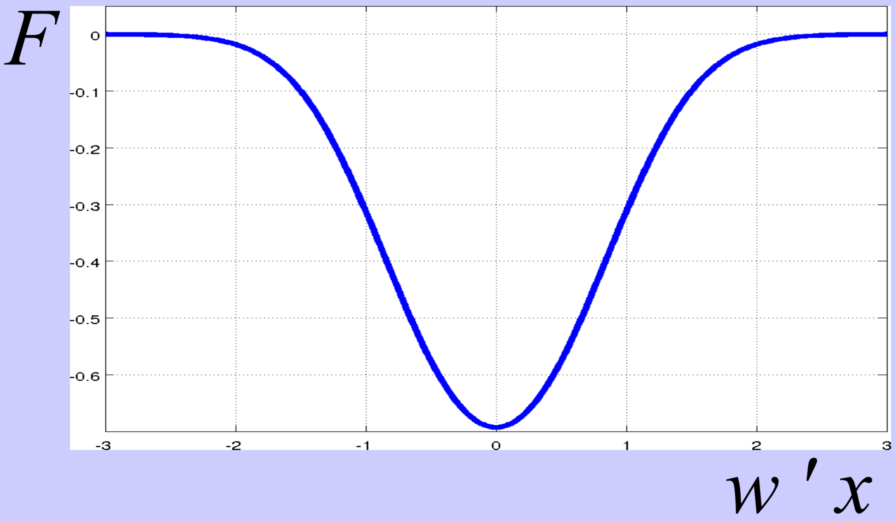
Interpretation



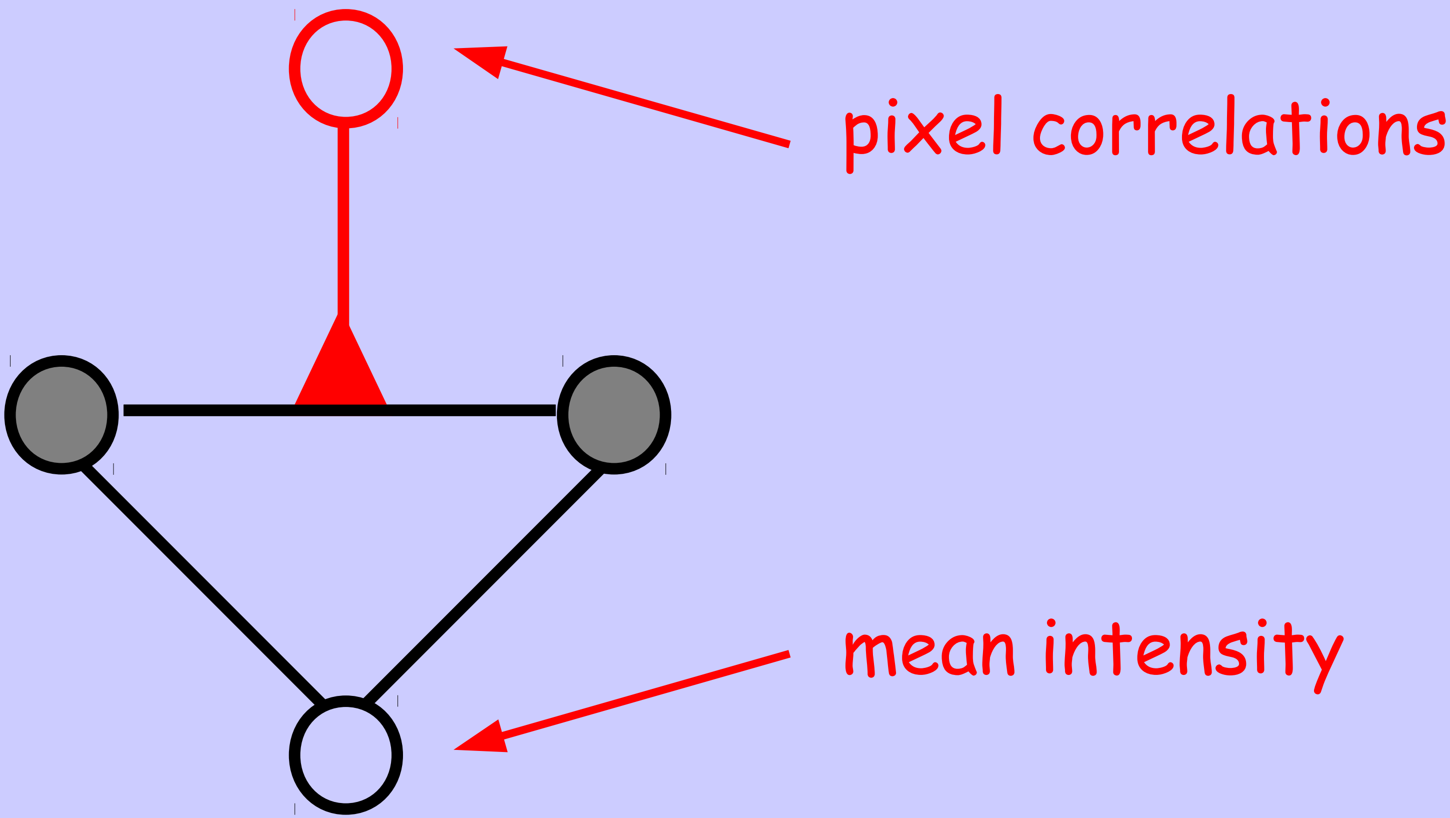
Integrating out latent variable, we get "robust" error metric.

$$F = -\log[e^{-0*(w'x)^2+b*0} + e^{-(w'x)^2+b}]$$

$$= -\log[1+e^{-(w'x)^2+b}]$$

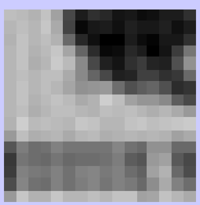


Interpretation

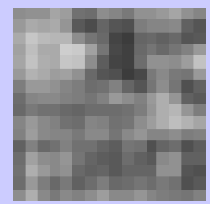


How mean & covariance units cooperate

input



reconstruction using only mean units



$$Wh^m$$

reconstruction using both mean&cov units

$$\Sigma(h^c)\cdot(Wh^m)$$

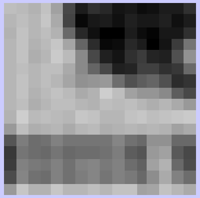
$$p(x|h^c, h^m) = N(\Sigma(Wh^m), \Sigma)$$

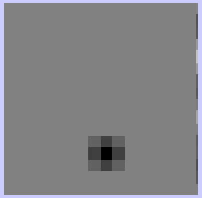
$$\Sigma^{-1} = C \, diag[P h^c] C' + I$$

How mean & covariance units cooperate

setting mean unit reconstruction by hand

input





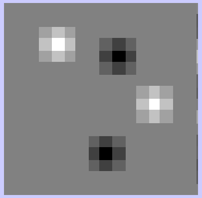
M

reconstruciton using covariance units

$\Sigma(h^c)\cdot M$

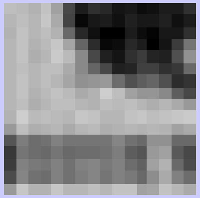
How mean & covariance units cooperate

setting mean unit reconstruction by hand

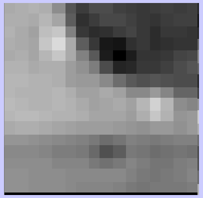


M

input



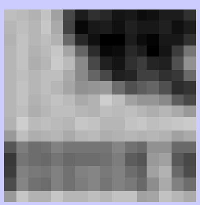
reconstruciton using covariance units



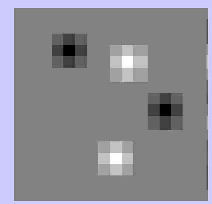
$\Sigma(h^c) \cdot M$

How mean & covariance units cooperate

input

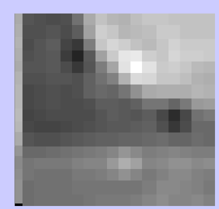


setting mean unit reconstruction by hand



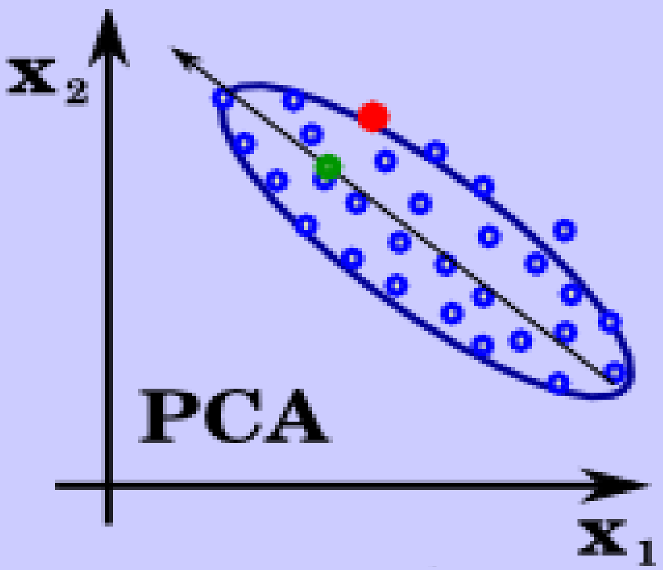
$$M$$

reconstruciton using covariance units

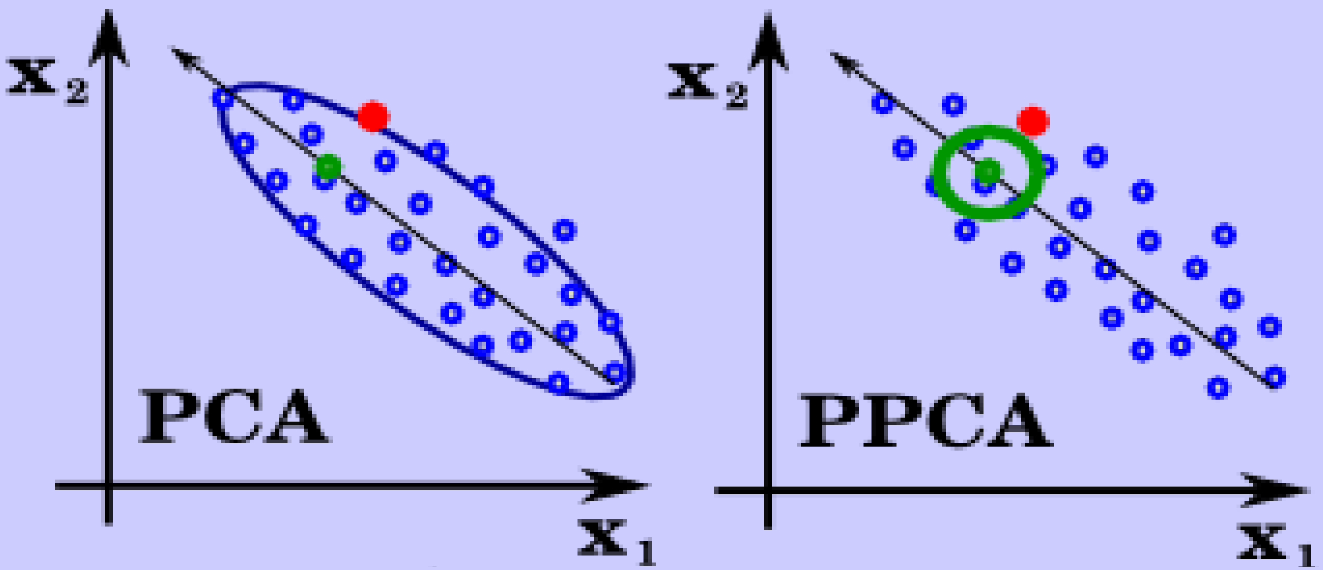


$$\Sigma\left(h^{c}\right)\cdot M$$

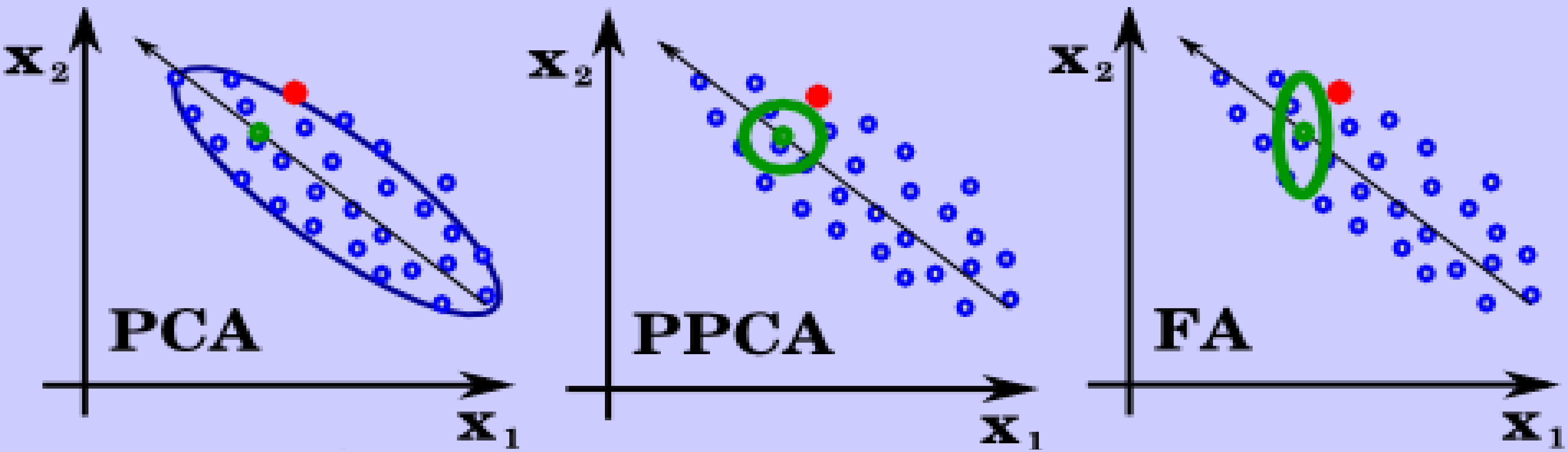
Comparison



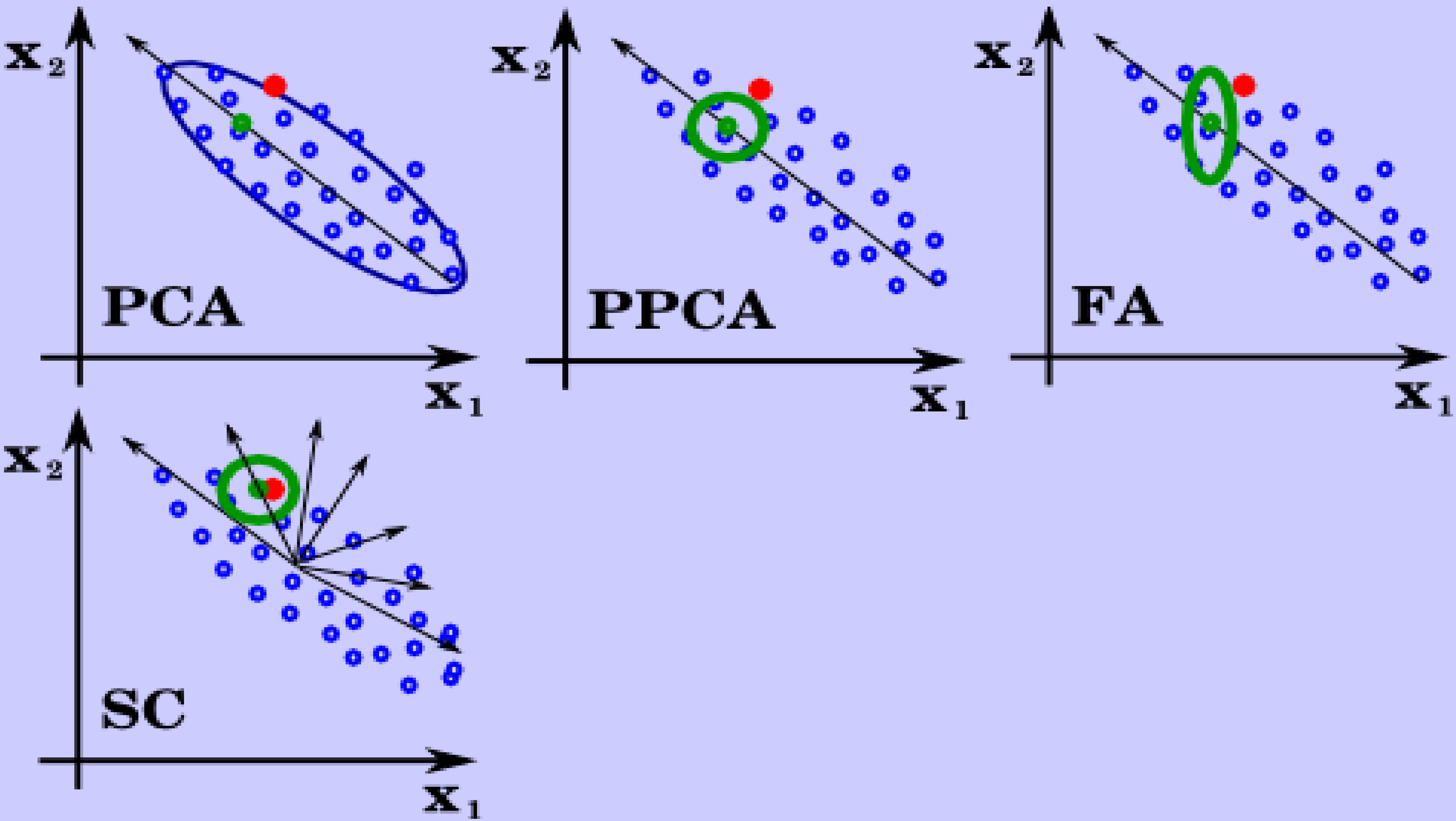
Comparison



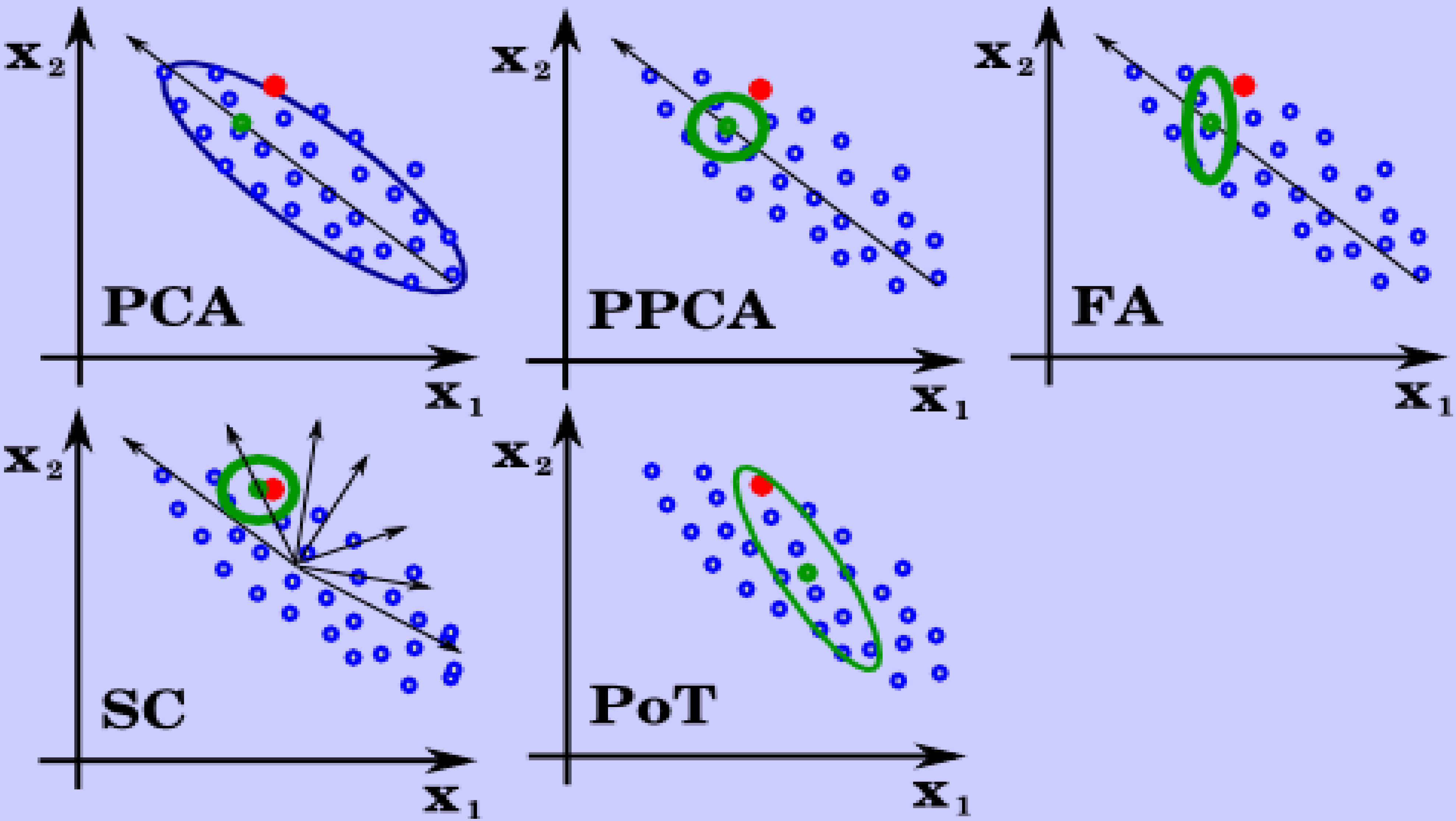
Comparison



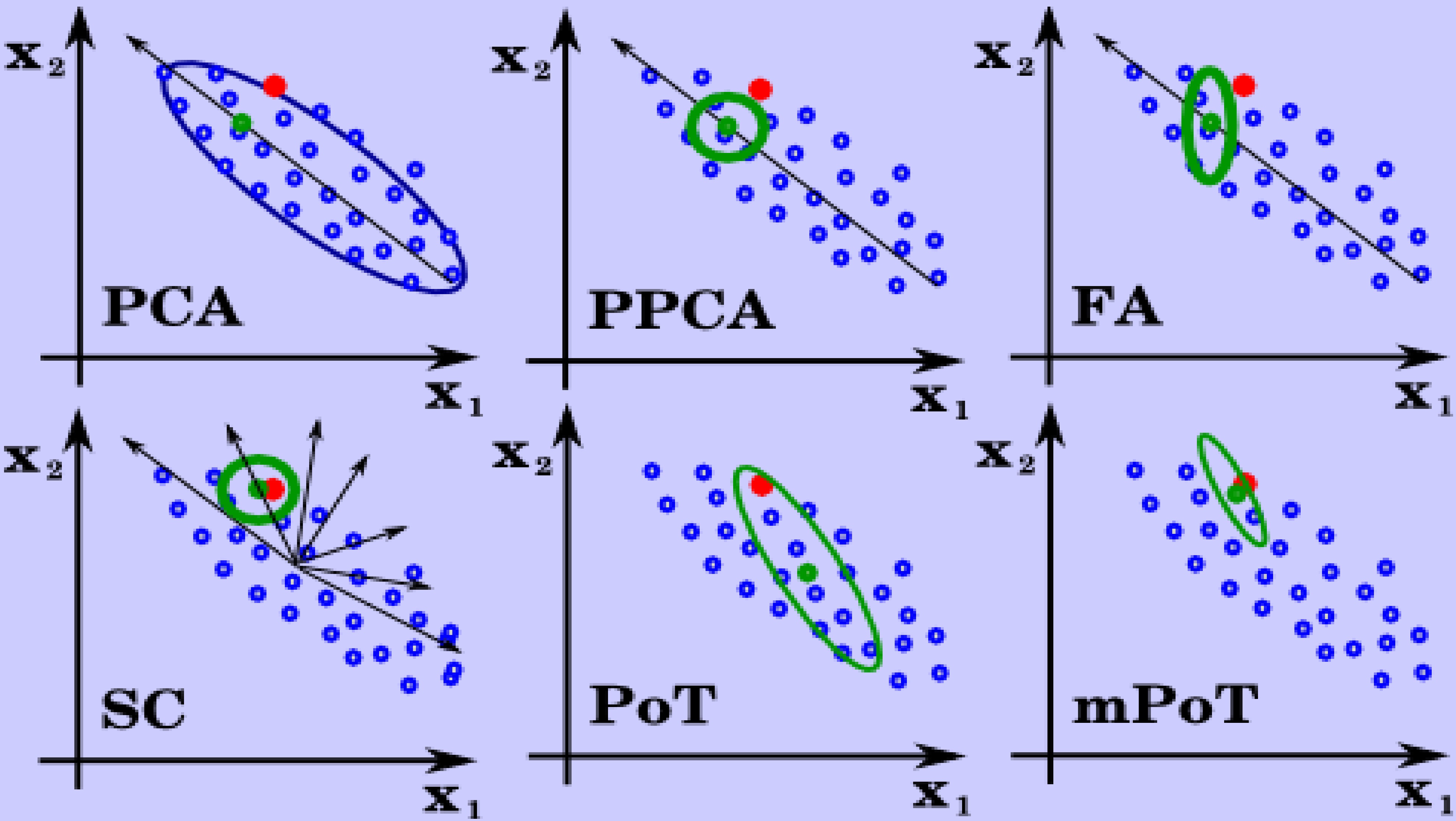
Comparison



Comparison

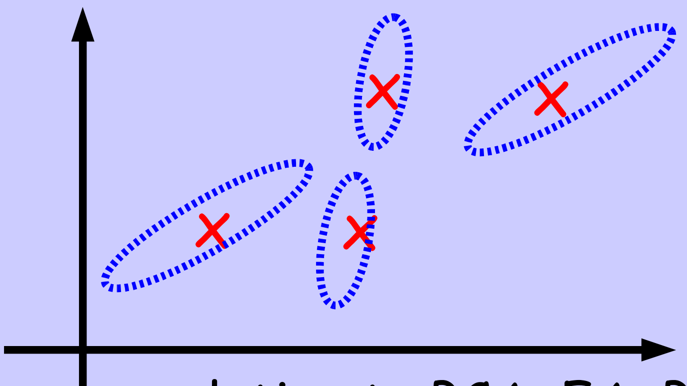


Comparison



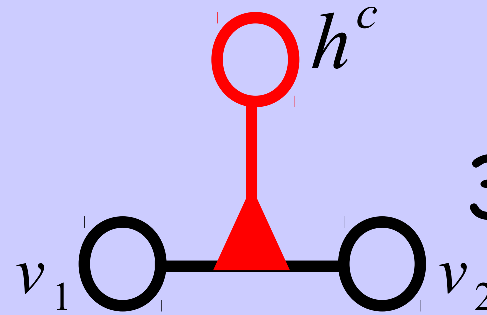
Relation to prior work

- Looking at $p(v|h)$



- relation to PCA, FA, PoT, etc.

- Looking at $E(v,h)$

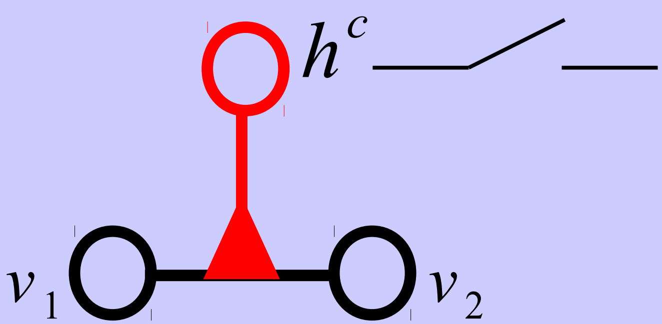


3rd order BM

- relation to conditional 3-way RBM

Memisevic et al 07, Taylor et al. 2009

- Looking at hiddens

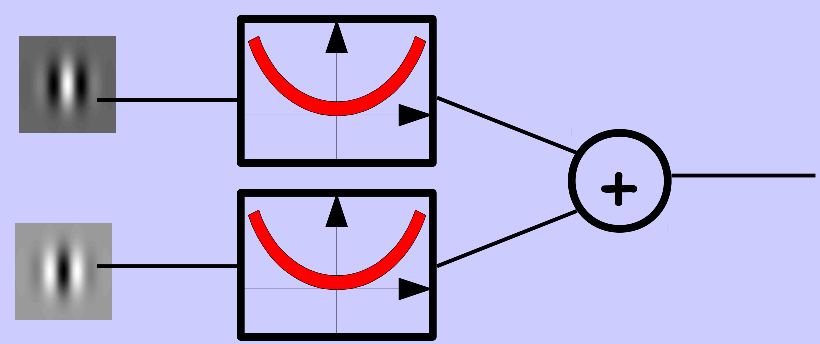


- relation to line process and PoT

Geman etal 84, Blake etal 87, Black etal 96

- Looking at $p(h|v)$

$$p(h_k^c = 1|v) = \sigma(-\frac{1}{2} P_k (C'v)^2 + b_k)$$



- relation to simple-complex cell model

Outline

- mathematical formulation of the model

- training

- generation of natural images

- recognition of facial expression under occlusion

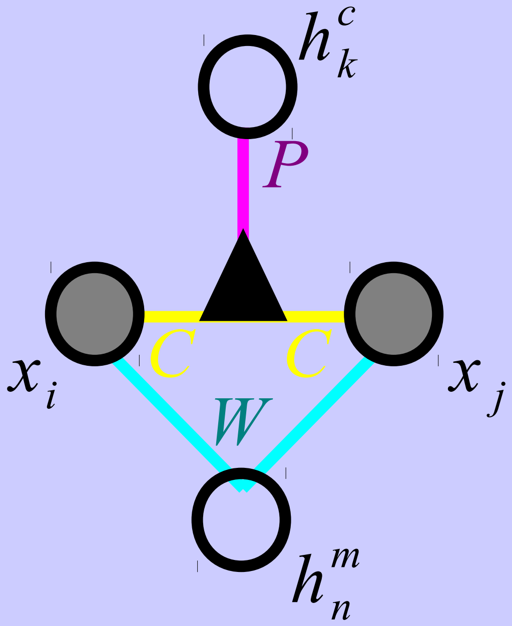
- learning acoustic features for spech recognition

- conclusion

Learning

- maximum likelihood $$p(x)=\frac{\int_{h^m,h^c} e^{-E(x,h^m,h^c)}}{\int_{x,h^m,h^c} e^{-E(x,h^m,h^c)}}$$

  - Fast Persistent Contrastive Divergence
  - Hybrid Monte Carlo to draw samples



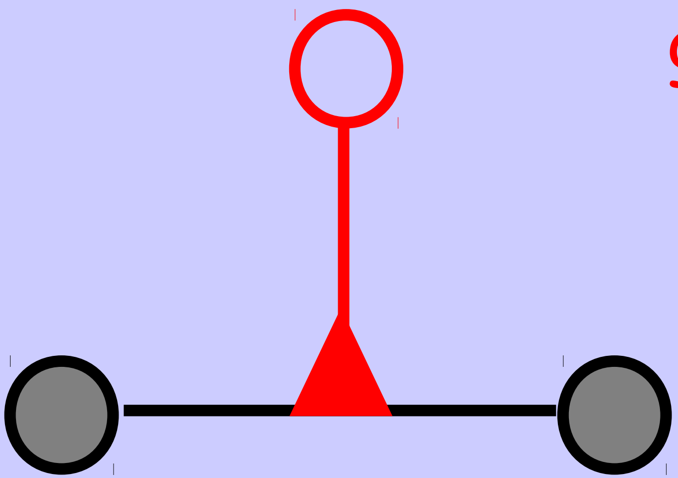
$$E=\frac{1}{2}x'C[diag(Ph^c)]C'x-x'Wh^m+\ldots$$

Learning

$$p(x)=\frac{\int_{h^m,h^c} e^{-E(x,h^m,h^c)}}{\int_{x,h^m,h^c} e^{-E(x,h^m,h^c)}}=\frac{e^{-F(x)}}{\int_x e^{-F(x)}}$$

$$F(x)=-\log\int_{h^m,h^c} e^{-E(x,h^m,h^c)}$$

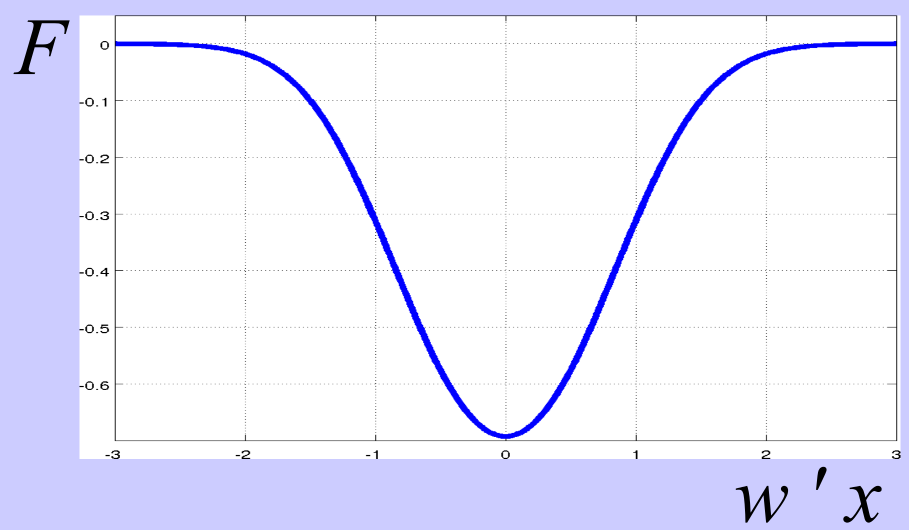
Interpretation



Integrating out latent variable, we get "robust" error metric.

$$F = -\log\left[e^{-0*(w'x)^2 + b*0} + e^{-(w'x)^2 + b}\right]$$

$$= -\log\left[1 + e^{-(w'x)^2 + b}\right]$$



Learning

$$p(x;\theta) = \frac{e^{-F(x;\theta)}}{\int_y e^{-F(y;\theta)}}$$

$$L(x;\theta) = -\log p(x;\theta)$$

$$\theta \leftarrow \theta - \eta \frac{\partial L}{\partial \theta}$$

Learning

$$p(x;\theta)=\frac{e^{-F(x;\theta)}}{\int_y e^{-F(y;\theta)}}$$

$$L(x;\theta)=-\log p(x;\theta)$$

$$\theta \leftarrow \theta-\eta\frac{\partial L}{\partial \theta}$$

$$\frac{\partial L}{\partial \theta}=<\frac{\partial F(x;\theta)}{\partial \theta}>_{x\sim TrainSet}-<\frac{\partial F(y;\theta)}{\partial \theta}>_{y\sim p(y;\theta)}$$

Learning

$$p(x;\theta)=\frac{e^{-F(x;\theta)}}{\int_y e^{-F(y;\theta)}}$$

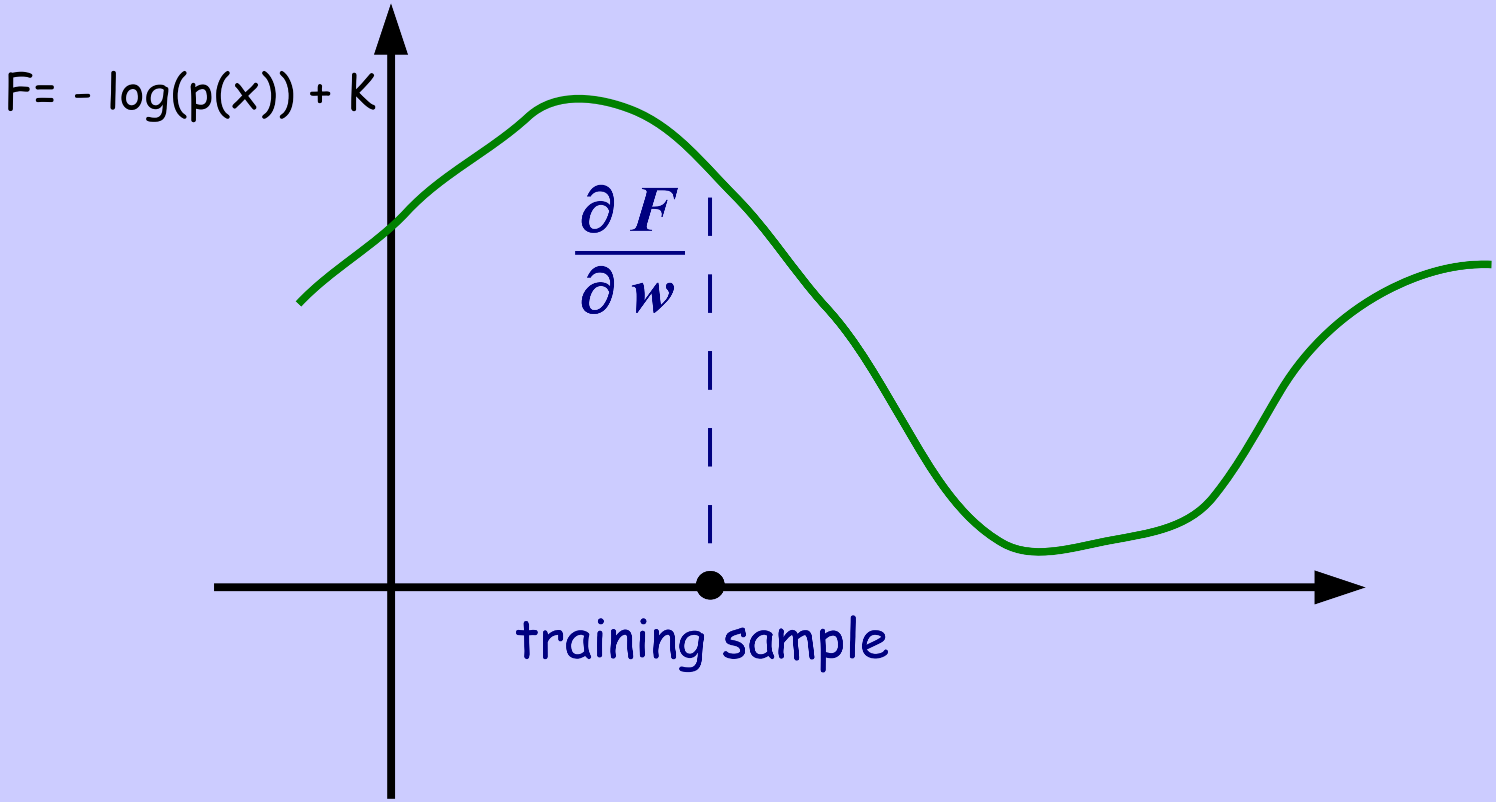
$$L(x;\theta)=-\log p(x;\theta)$$

$$\theta\leftarrow\theta-\eta\frac{\partial L}{\partial\theta}$$

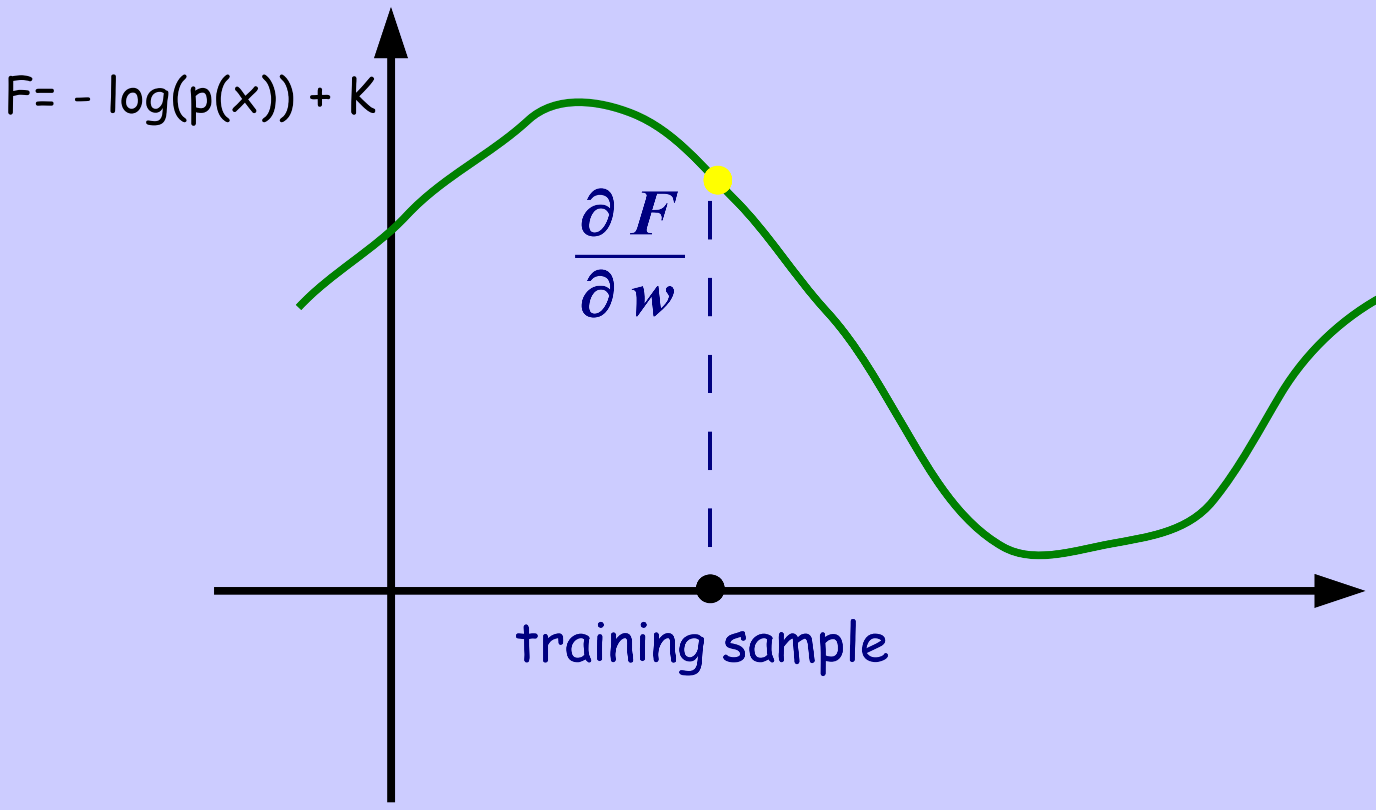
$$\frac{\partial L}{\partial\theta}=<\frac{\partial F(x;\theta)}{\partial\theta}>_{x\sim TrainSet}-<\frac{\partial F(y;\theta)}{\partial\theta}>_{y\sim p(y;\theta)}$$

We estimate this by using an MCMC method: HMC

Learning



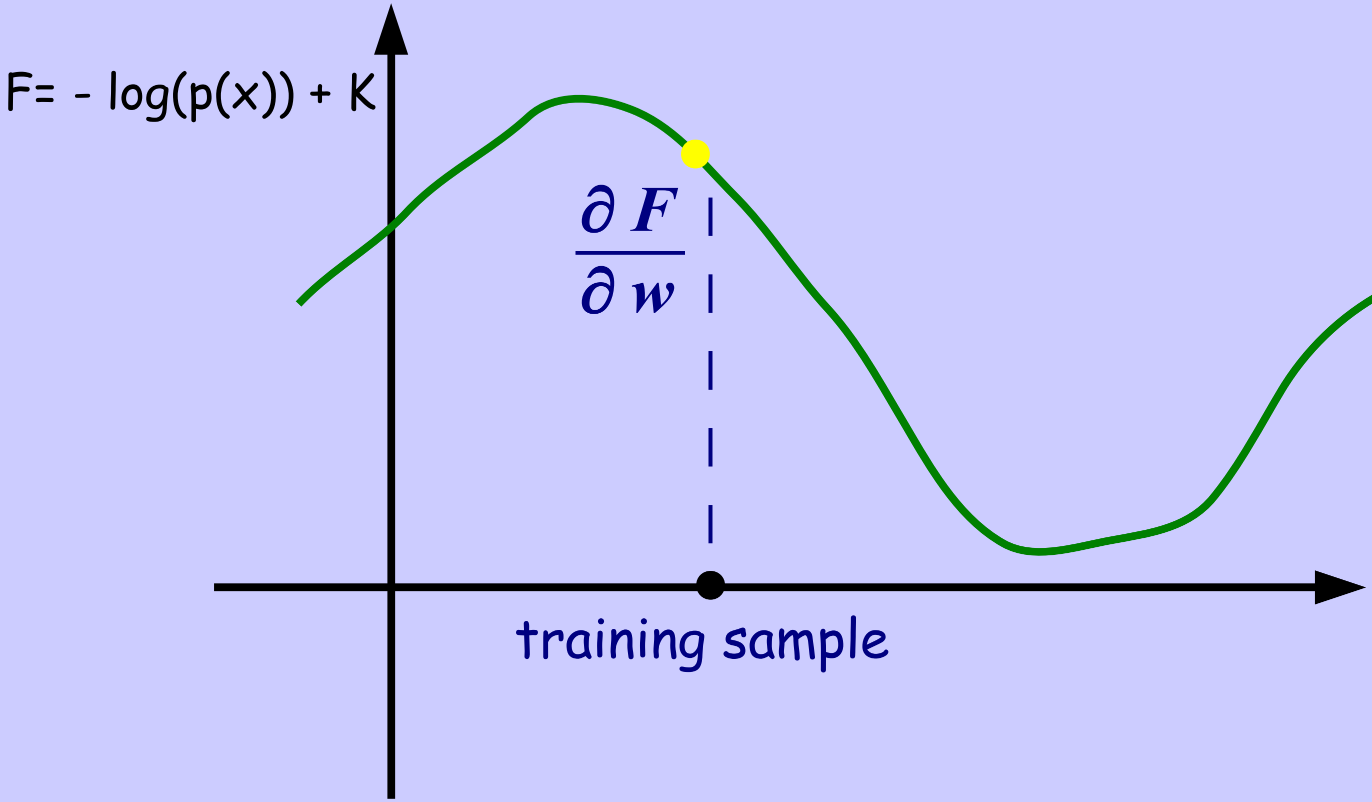
Learning



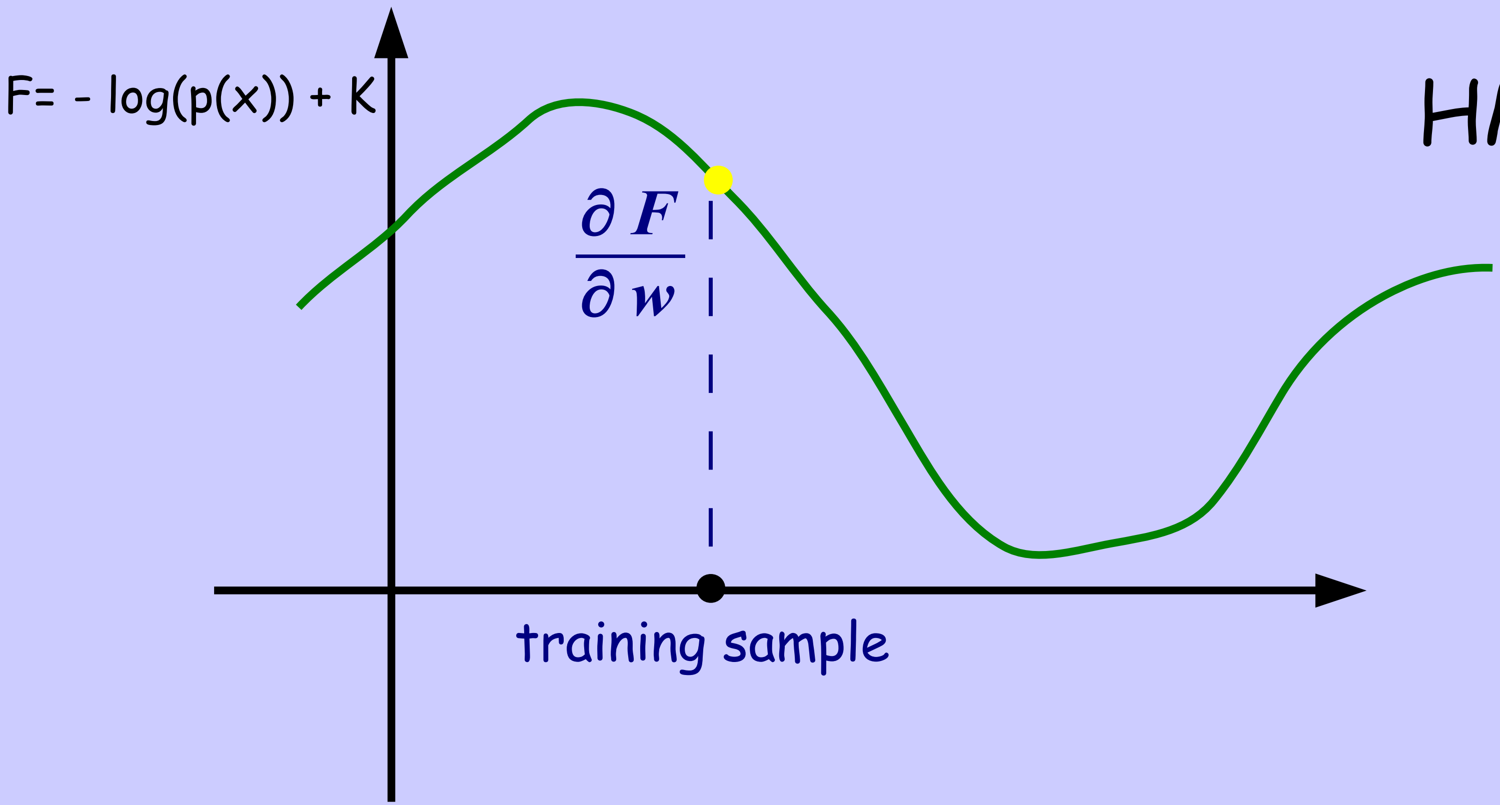
HMC dynamics

Learning

HMC dynamics

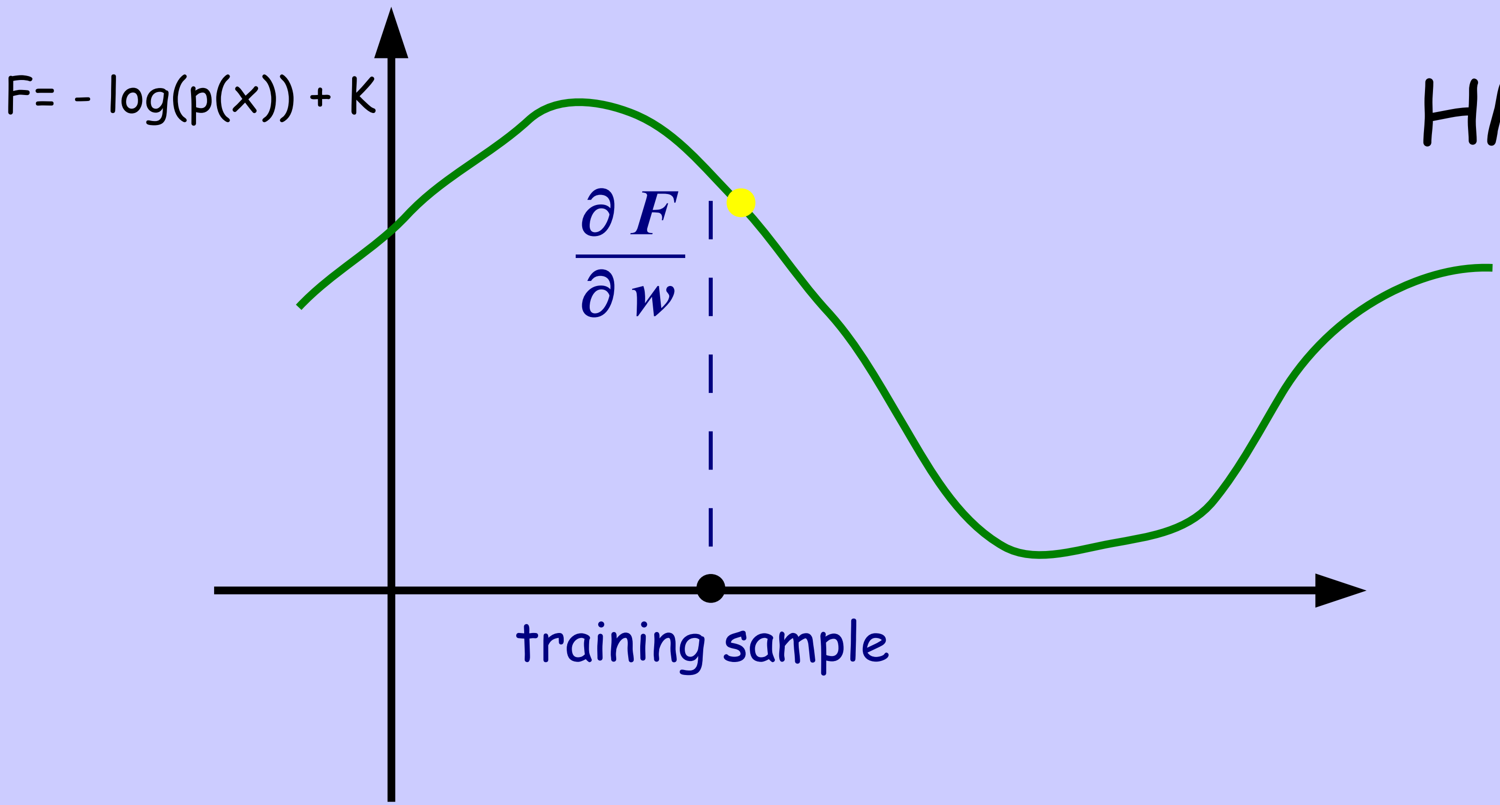


Learning



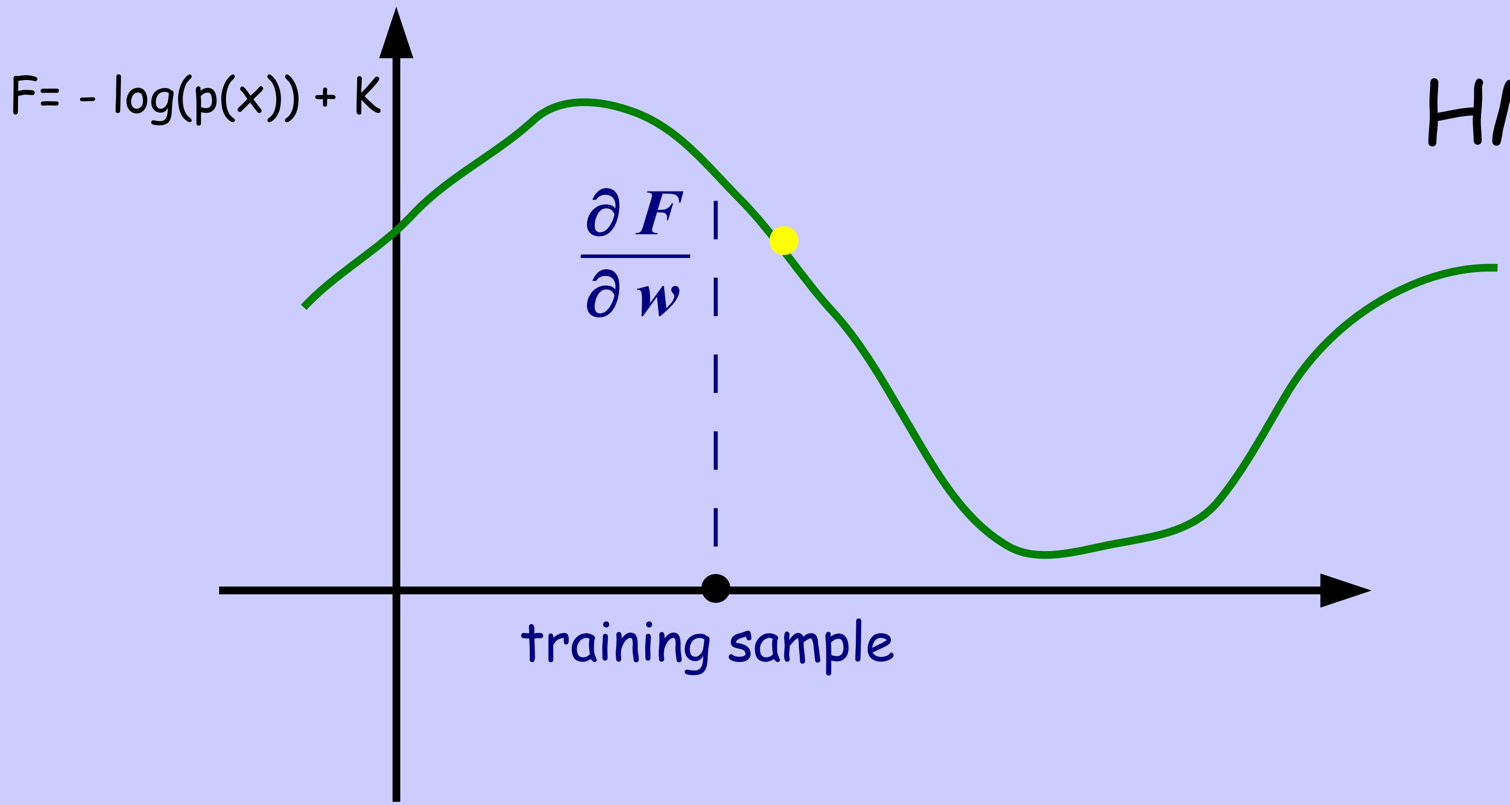
HMC dynamics

Learning



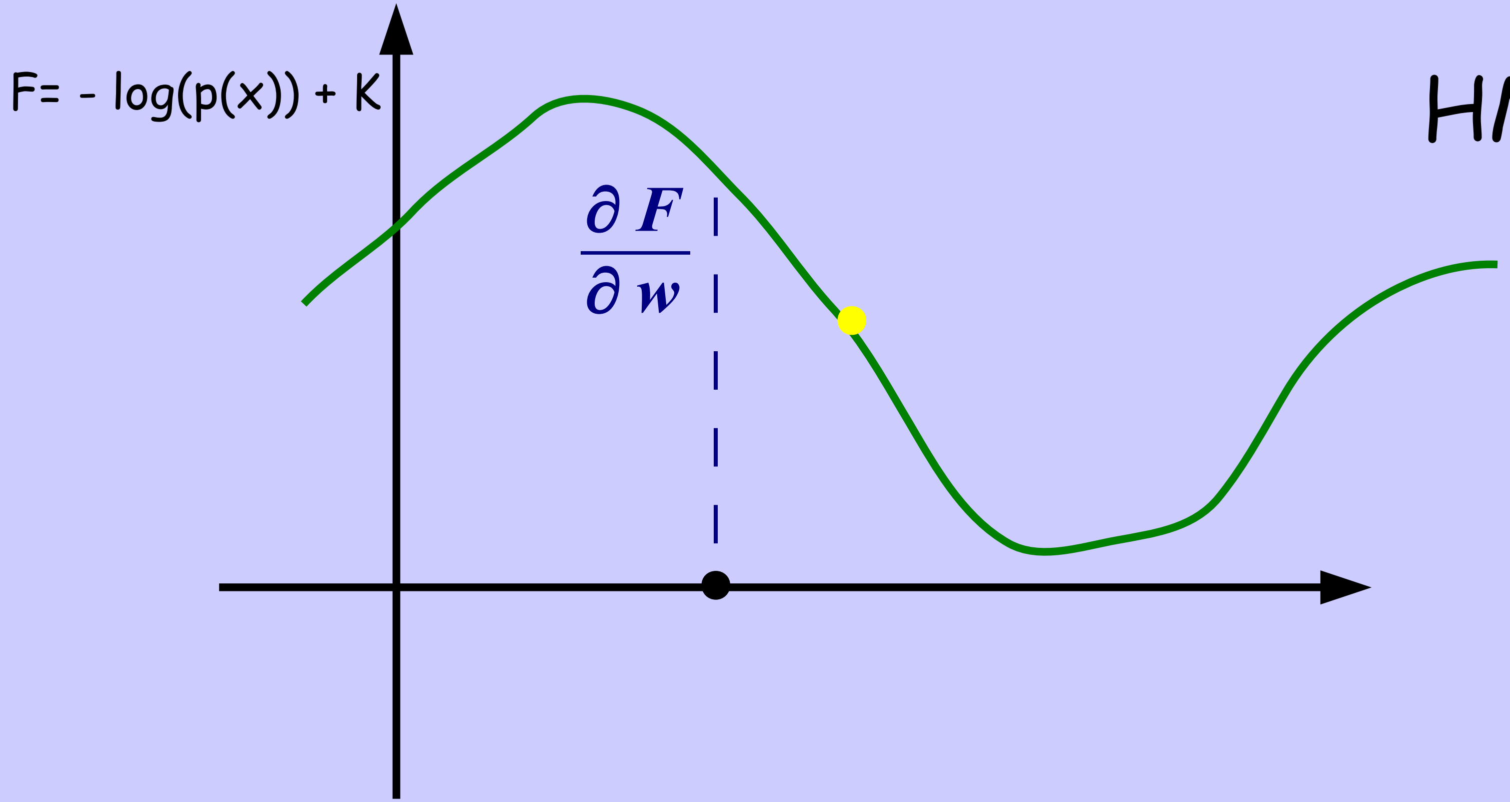
HMC dynamics

Learning



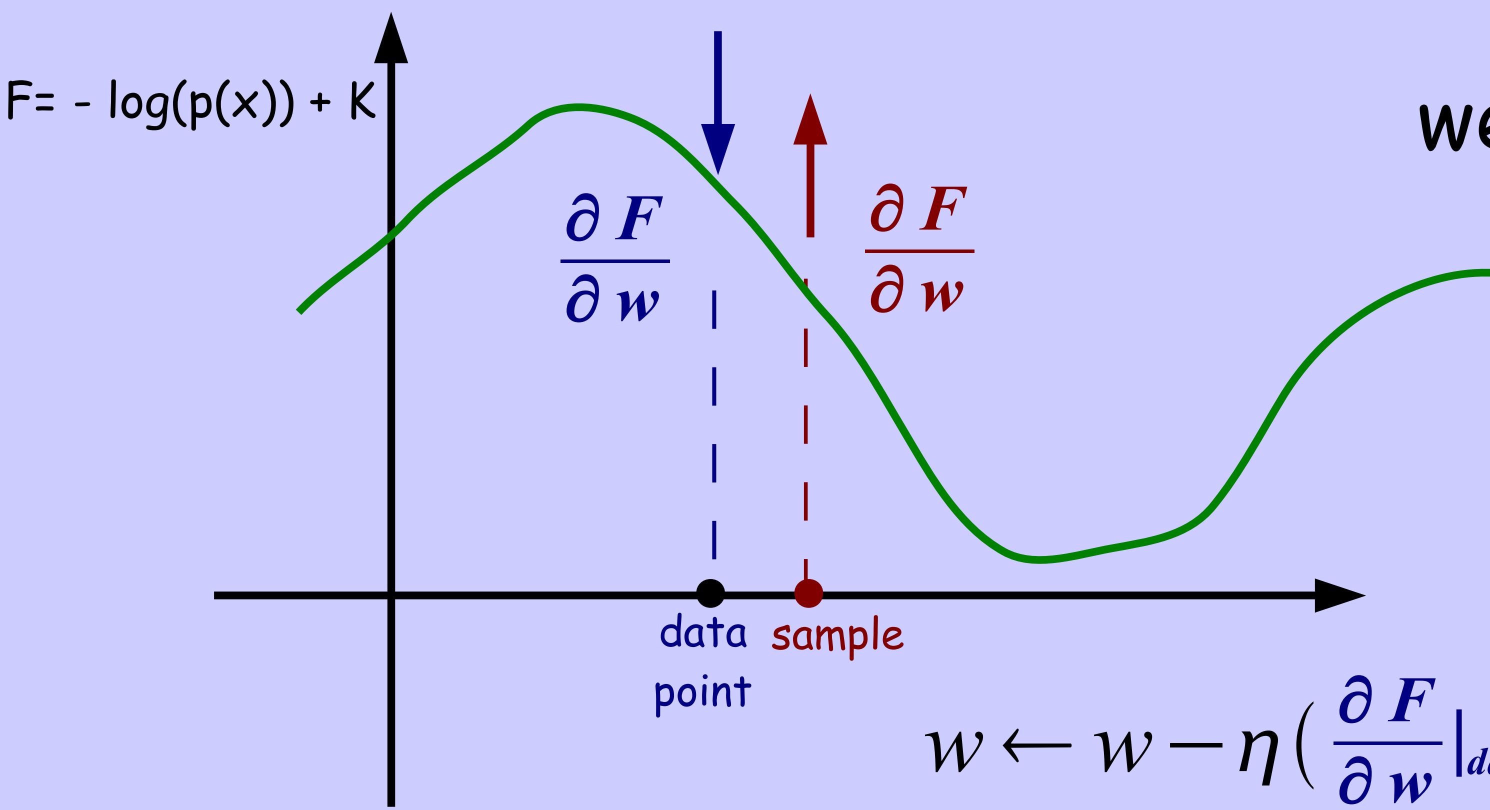
HMC dynamics

Learning



HMC dynamics

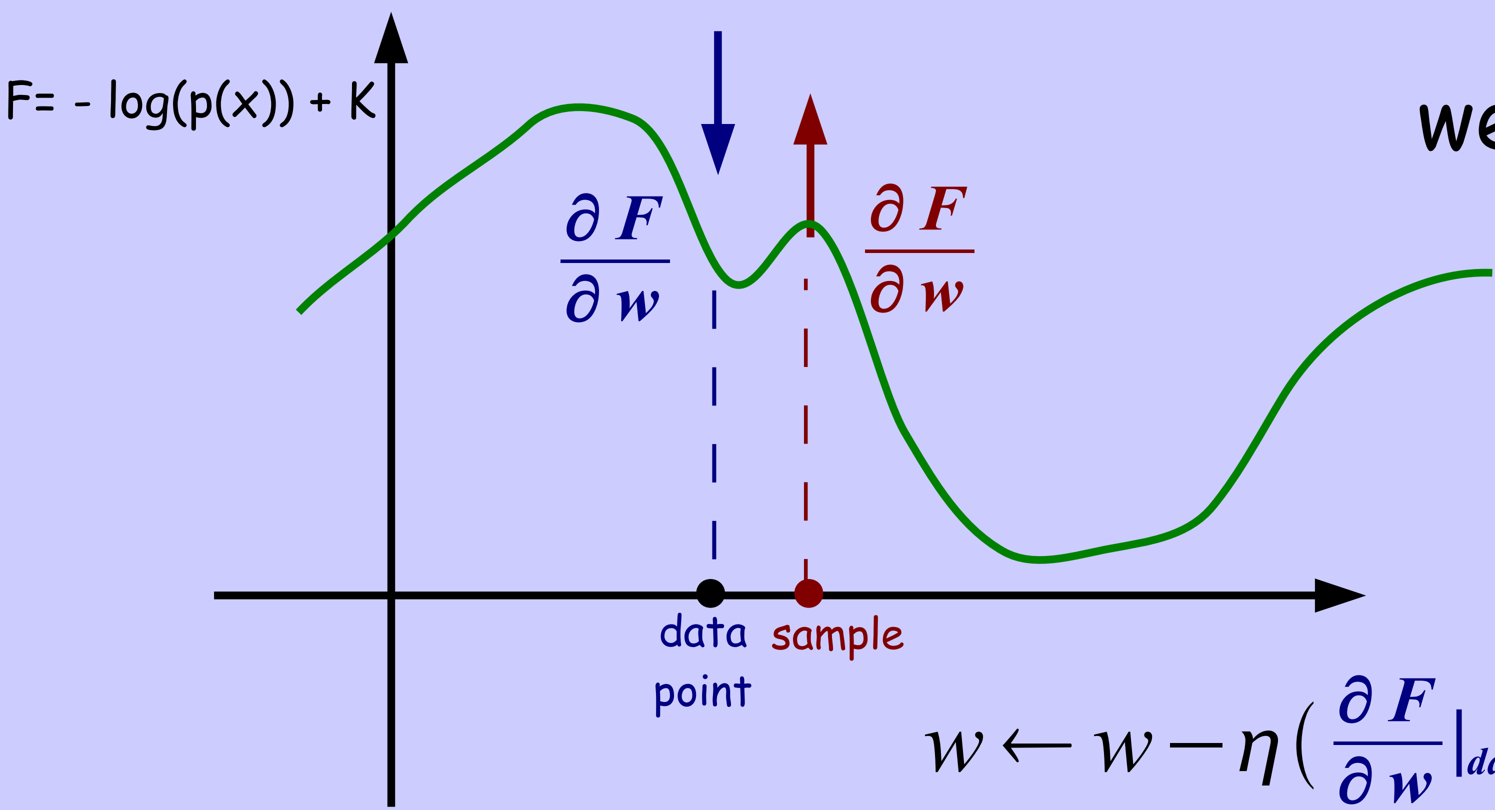
Learning



weight update

$$w \leftarrow w - \eta\left(\frac{\partial F}{\partial w}\Big|_{data} - \frac{\partial F}{\partial w}\Big|_{sample}\right)$$

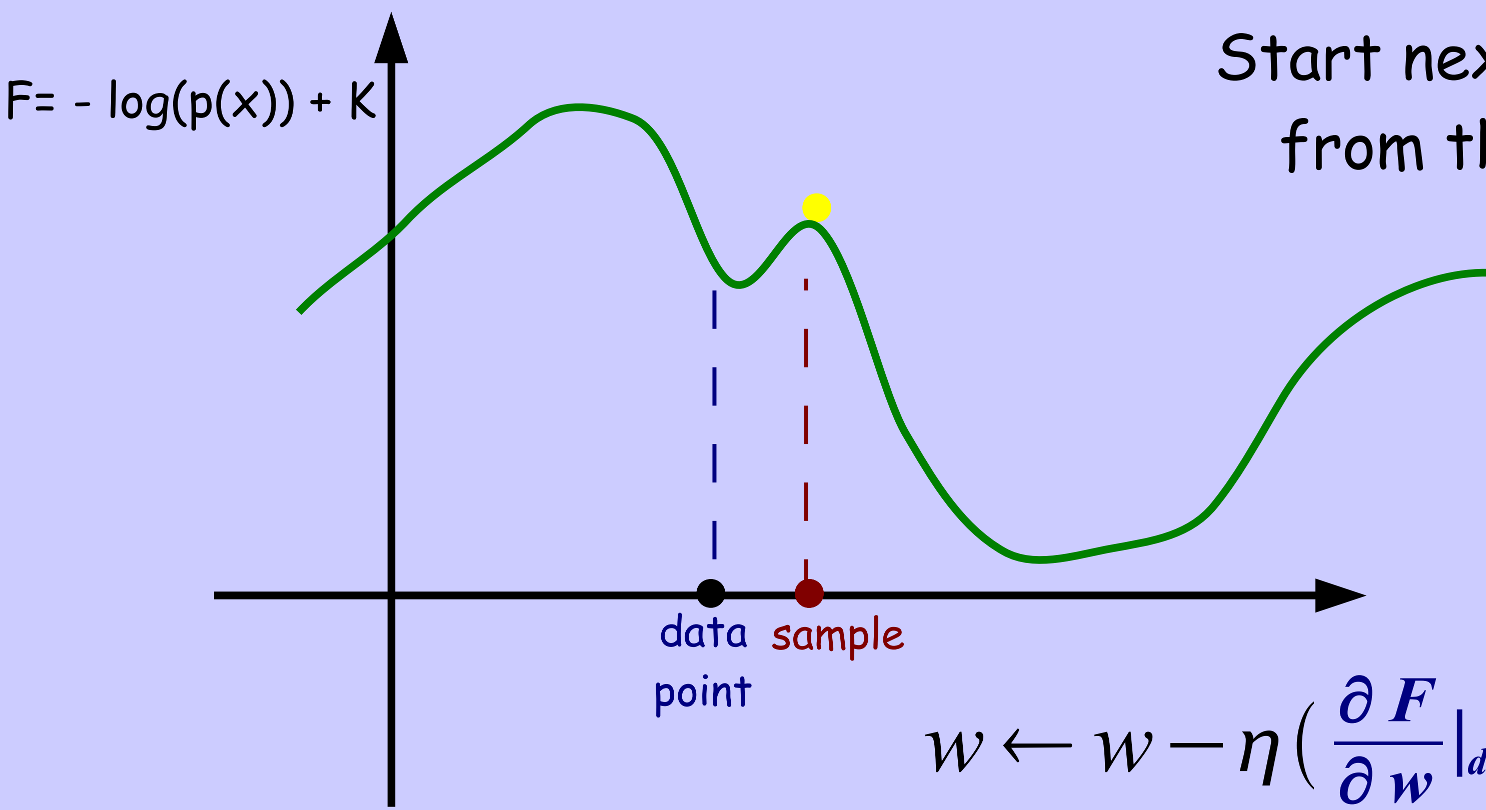
Learning



weight update

$$w \leftarrow w - \eta \left(\frac{\partial F}{\partial w}\Big|_{data} - \frac{\partial F}{\partial w}\Big|_{sample} \right)$$

Learning



Start next dynamics
from this sample

$$w \leftarrow w - \eta \left(\frac{\partial F}{\partial w}\Big|_{data} - \frac{\partial F}{\partial w}\Big|_{sample} \right)$$

Outline

- mathematical formulation of the model

- training

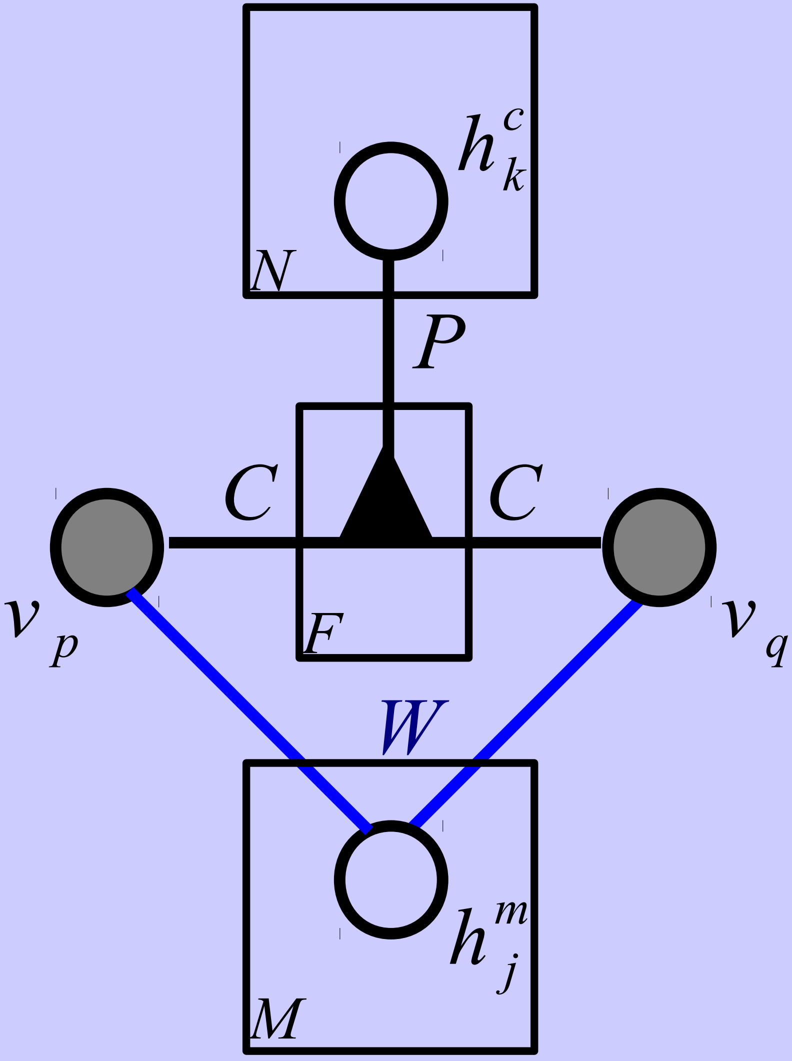
- generation of natural images

- recognition of facial expression under occlusion

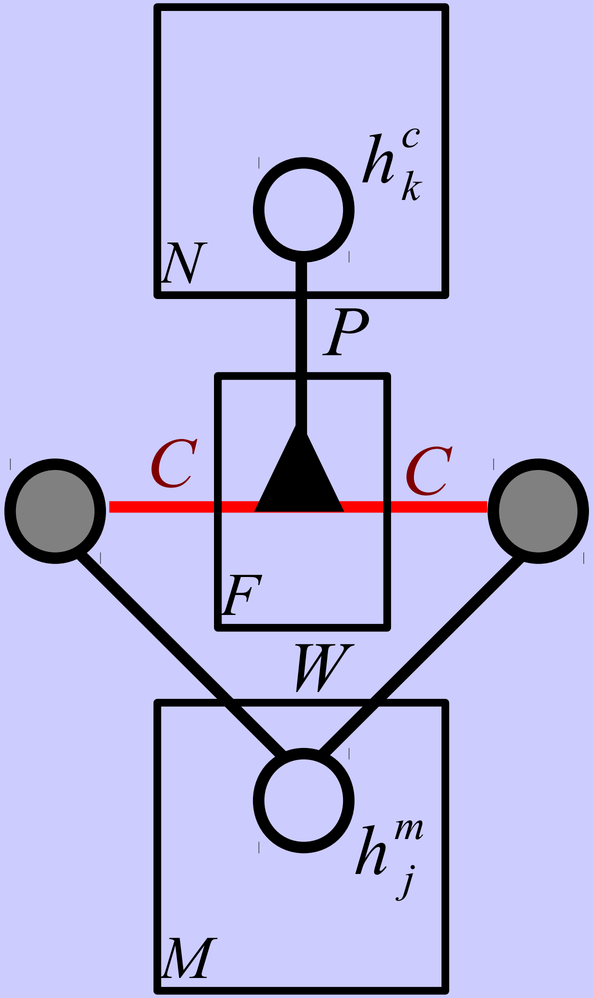
- learning acoustic features for spech recognition

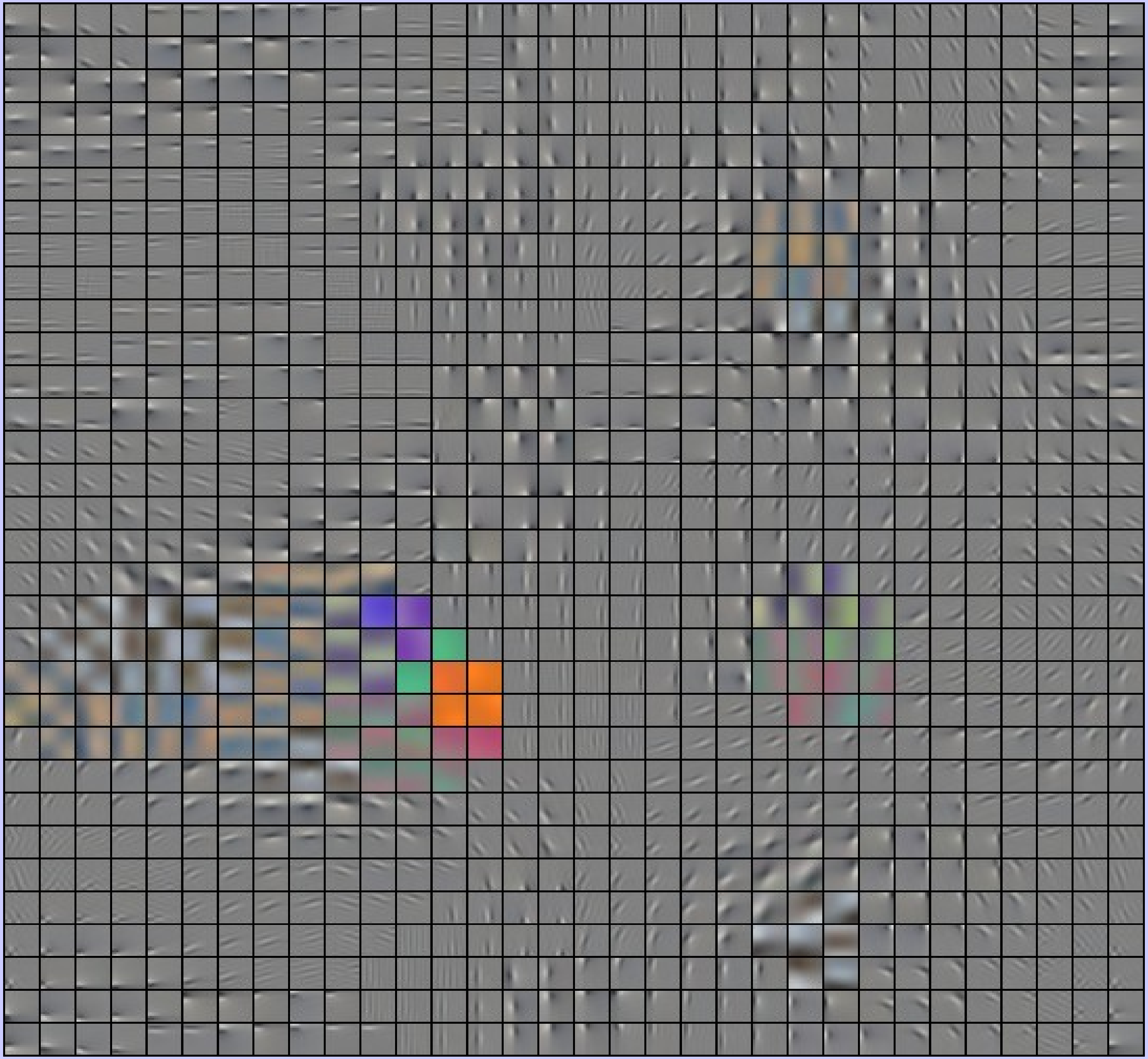
- conclusion

Learned Filters: mean filters W

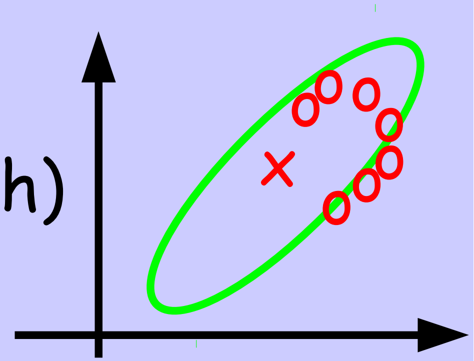


Learned Filters: covariance filters C



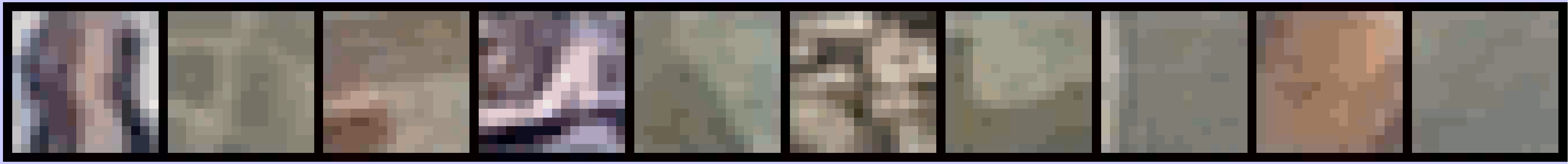


Random Walk: p(v|h)



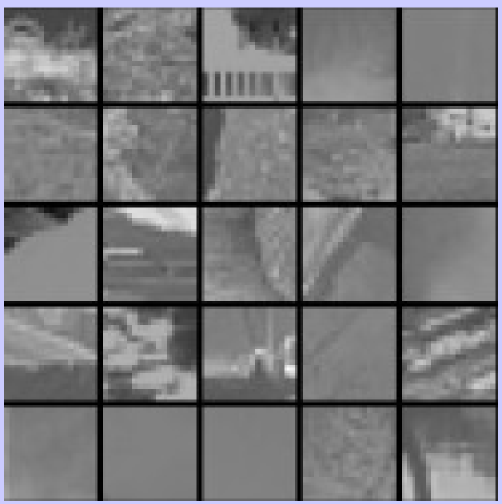
1) given image -> infer latent variables using p(h|v)

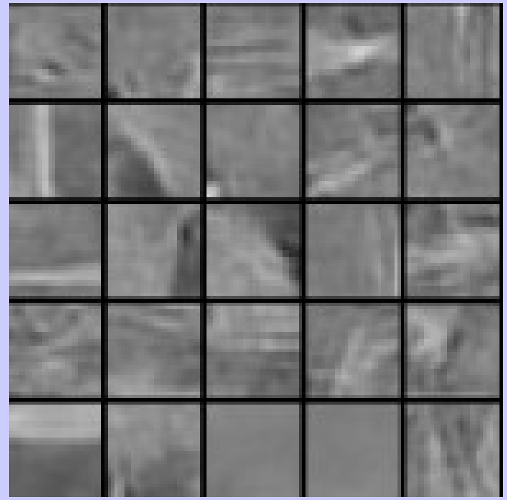
2) keeping latent variables fixed, sample from p(v|h)



Generation natural image patches

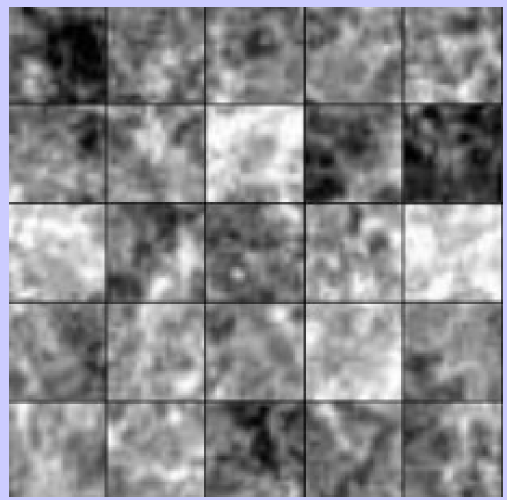
Natural images





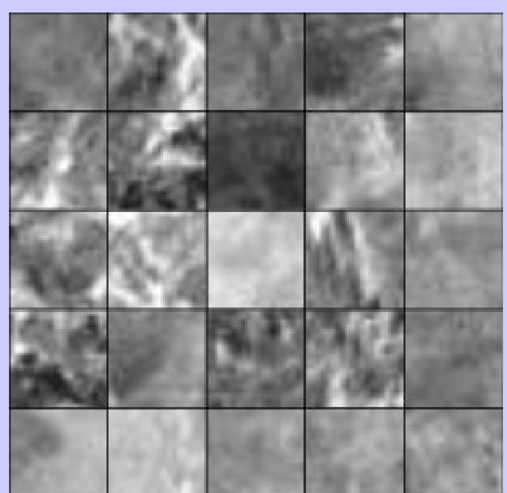
mcRBM

Ranzato and Hinton CVPR 2010



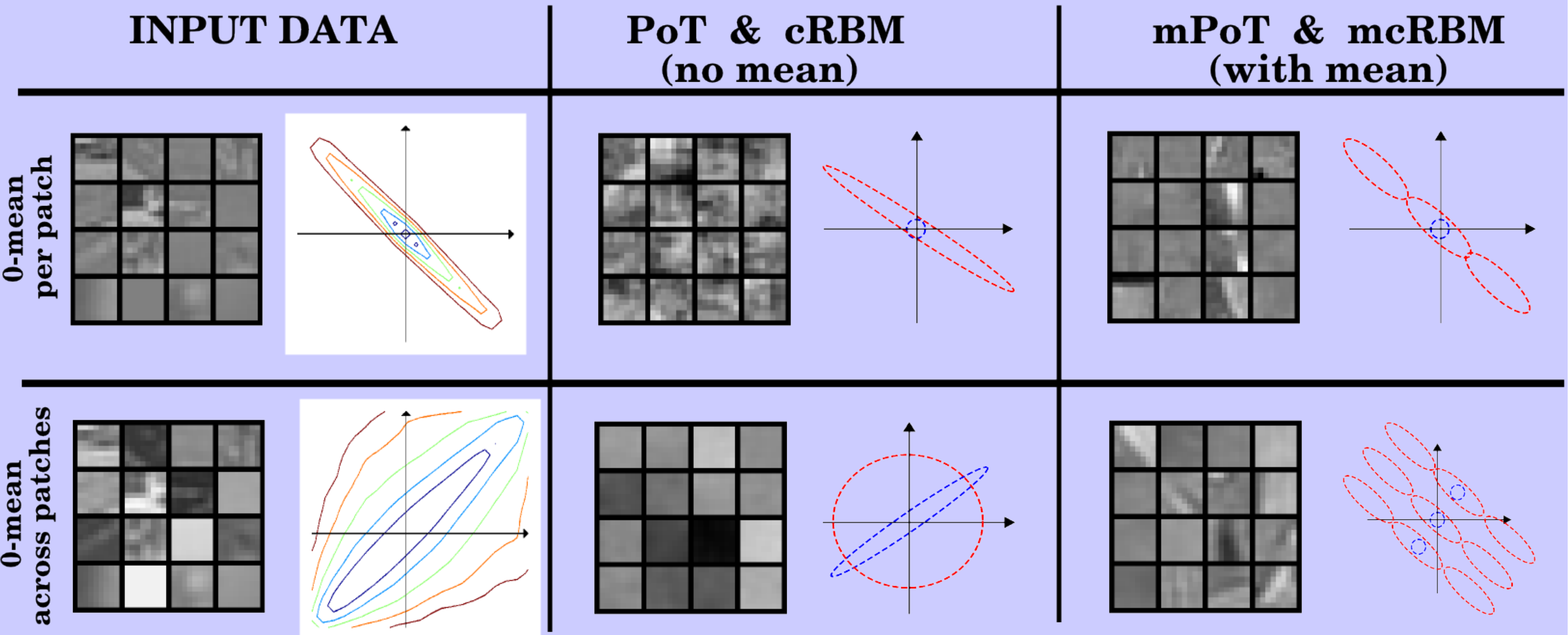
GRBM

from Osindero and Hinton NIPS 2008

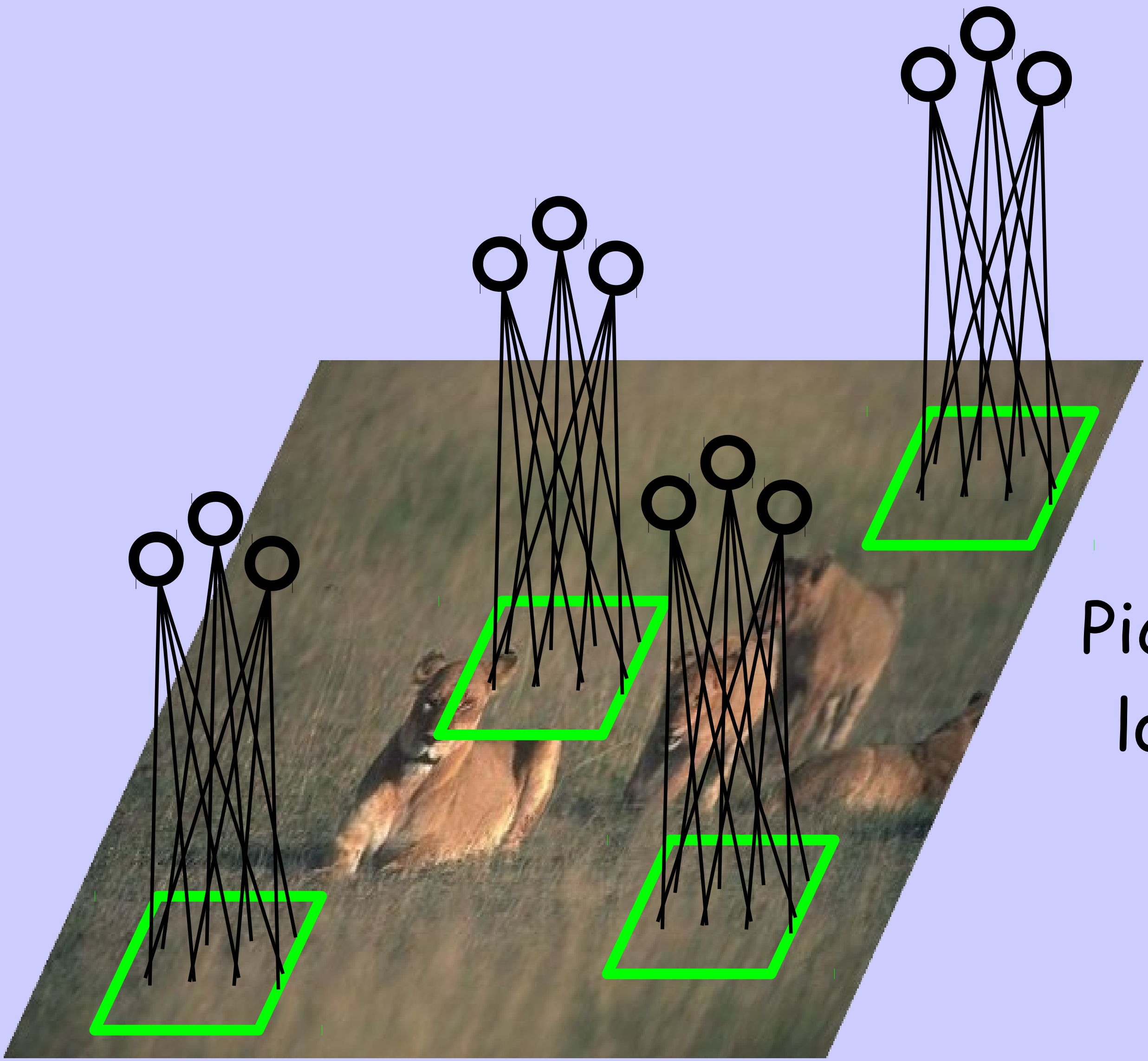


S-RBM + DBN

from Osindero and Hinton NIPS 2008


INPUT DATA
PoT & cRBM
(no mean)
mPoT & mcRBM
(with mean)
0-mean
per patch
0-mean
across patches

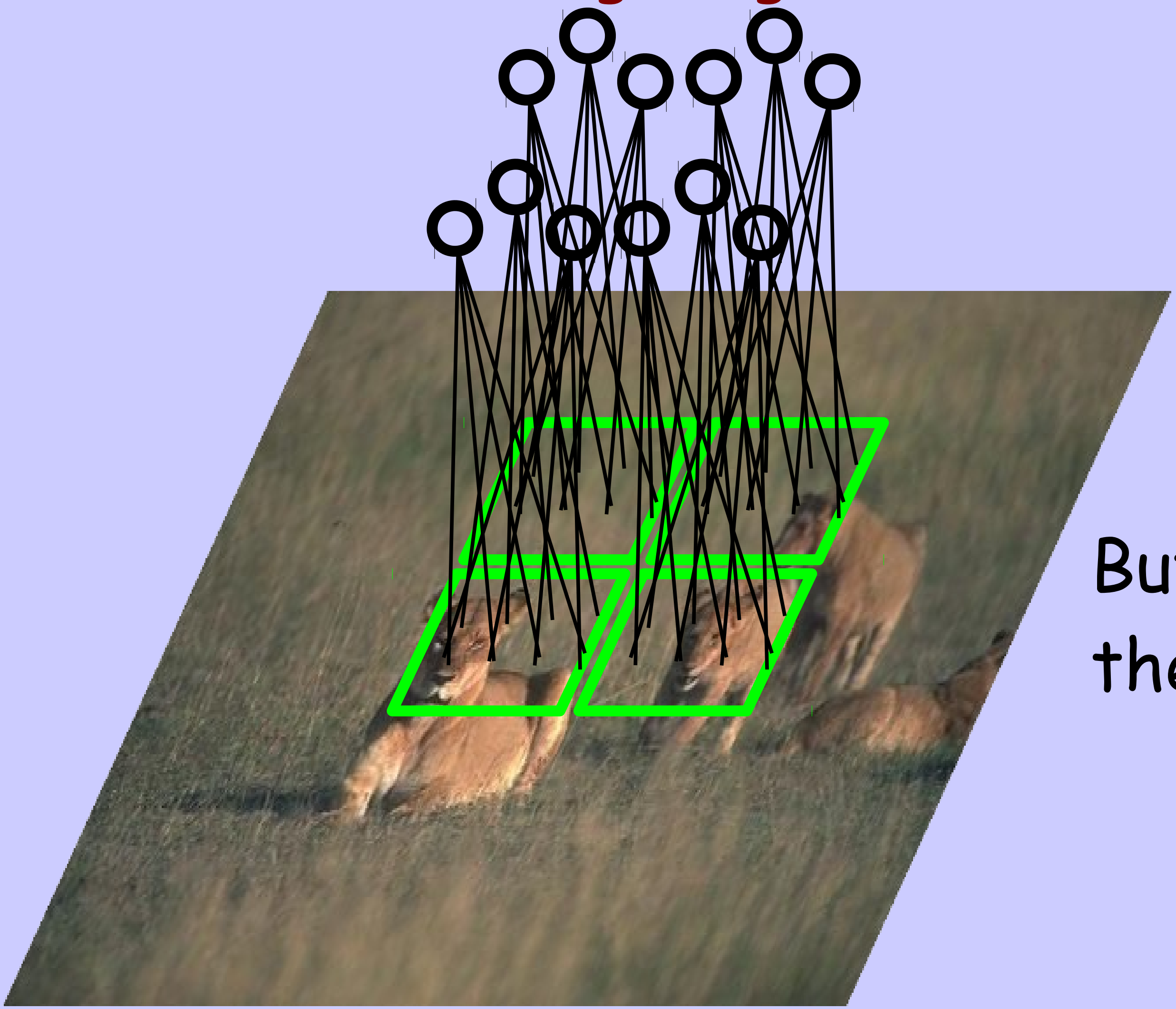
Training on Small Image Patches



Pick patches at random locations for training

From Patches to High-Resolution Images

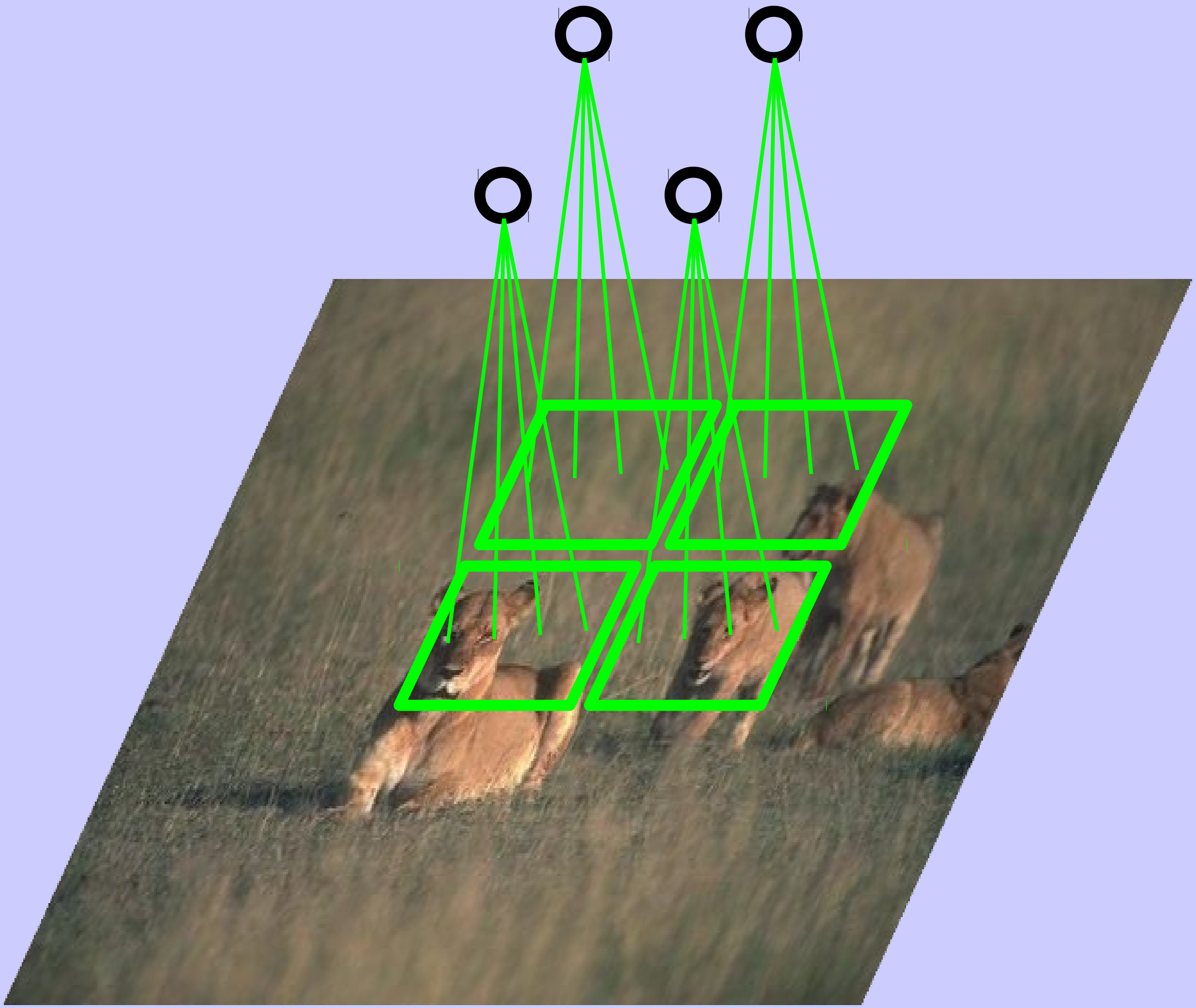
This is not a good way to extend the model to big images: block artifacts



But we could also take them from a grid

From Patches to High-Resolution Images

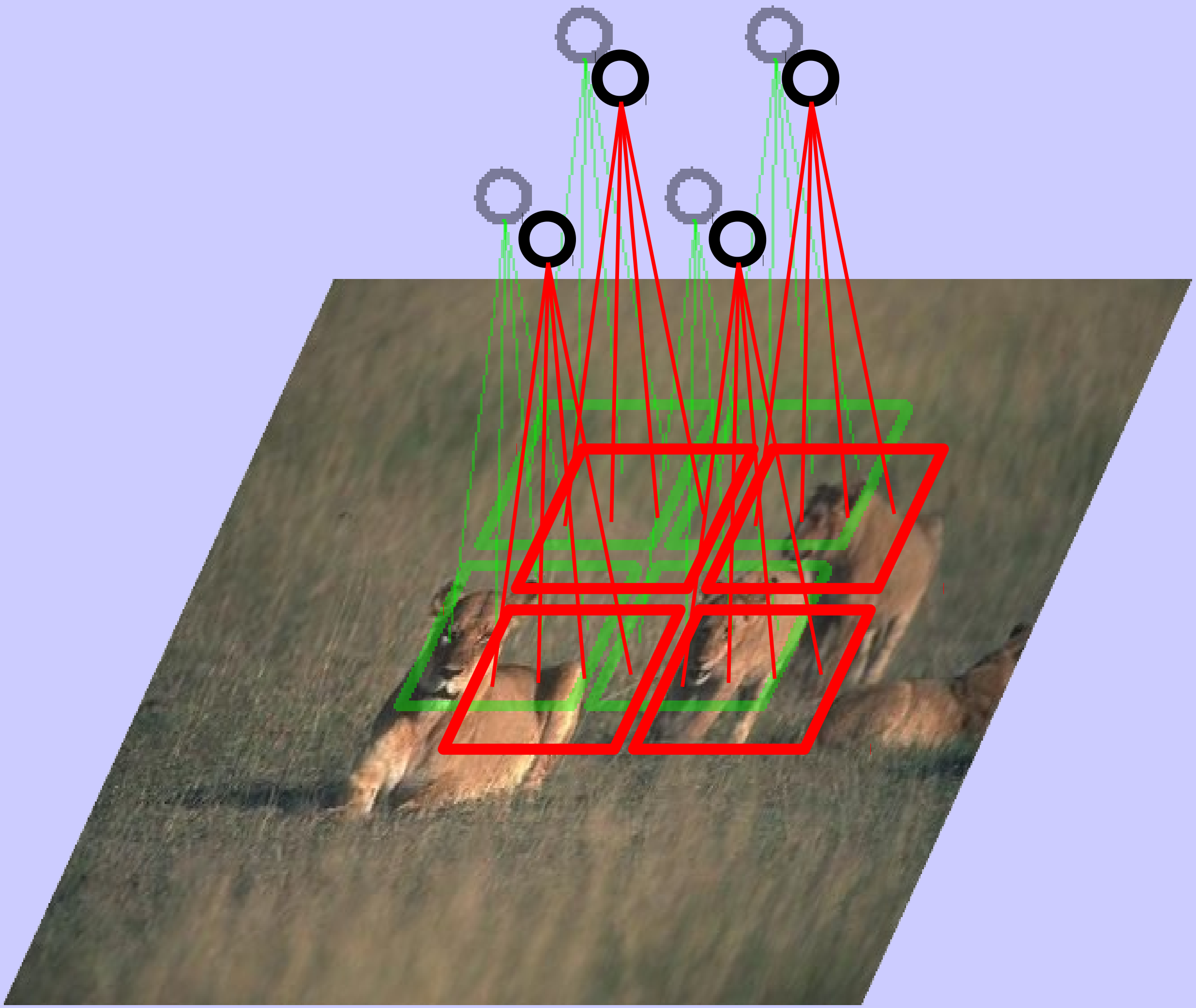
IDEA: have one subset of filters applied to these locations,



From Patches to High-Resolution Images

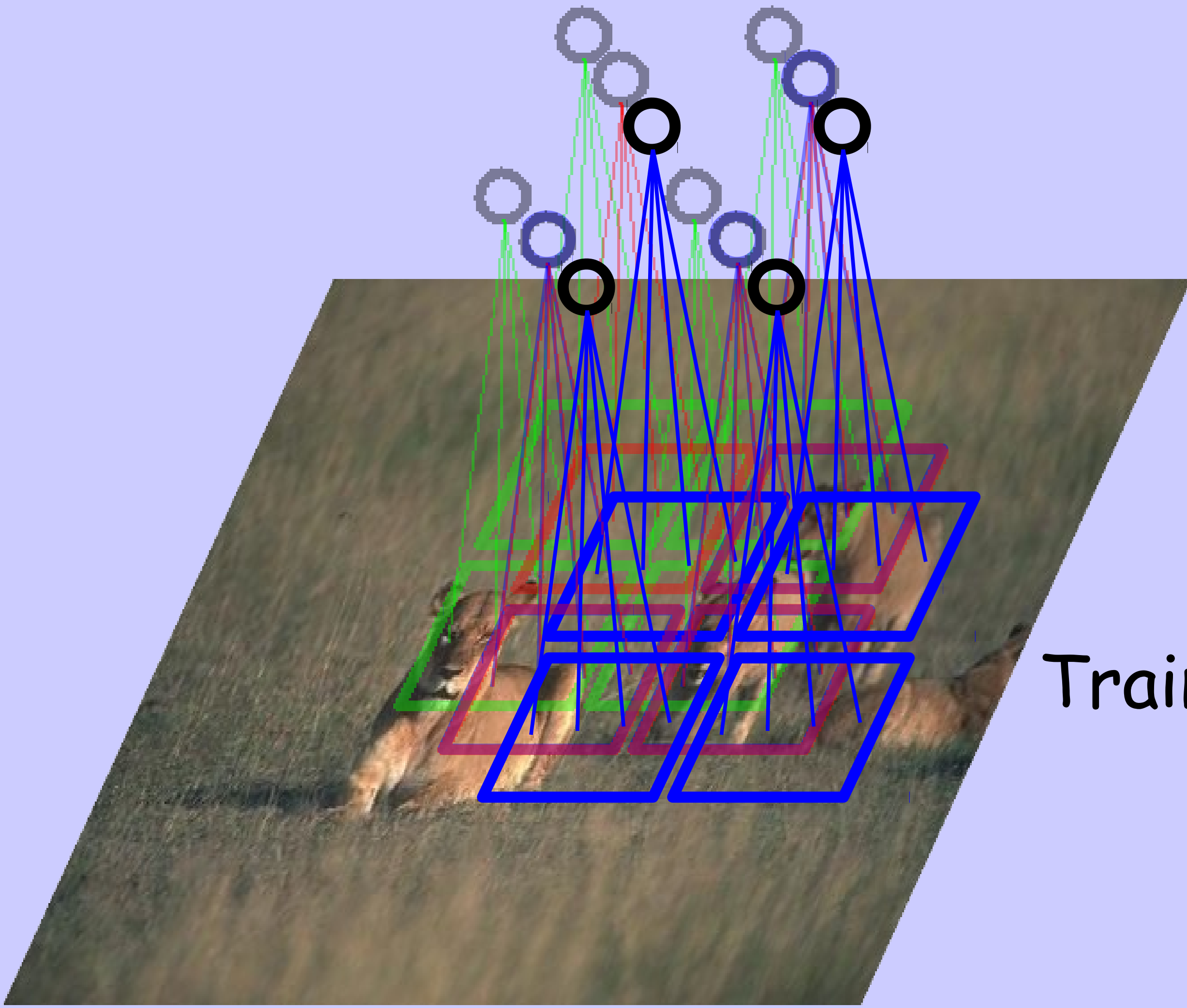
IDEA: have one subset of filters applied to these locations,

another subset to these locations



From Patches to High-Resolution Images

IDEA: have one subset of filters applied to these locations, another subset to these locations, etc.



Train jointly all parameters.

No block artifacts

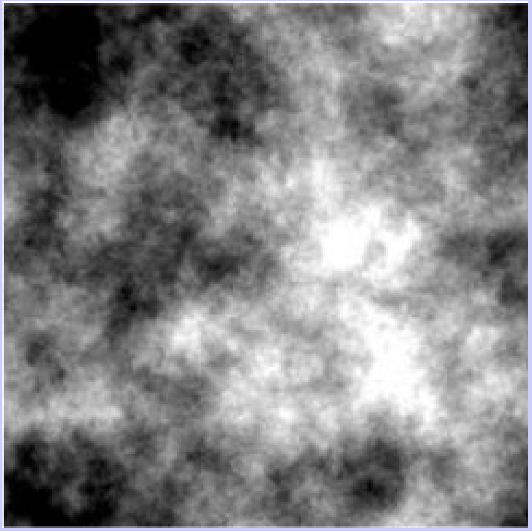
Reduced redundancy

Gregor LeCun arXiv 2010

Ranzato, Mnih, Hinton NIPS 2010

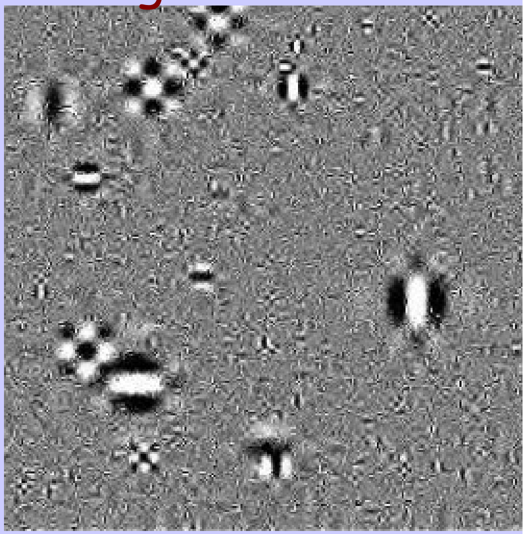
Sampling High-Resolution Images

Gaussian model



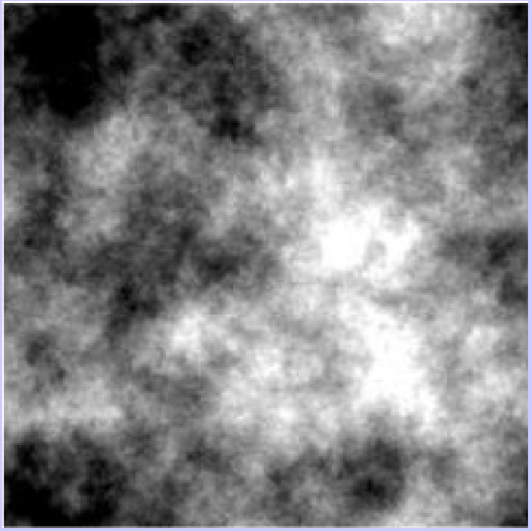
from Simoncelli 2005

marginal wavelet

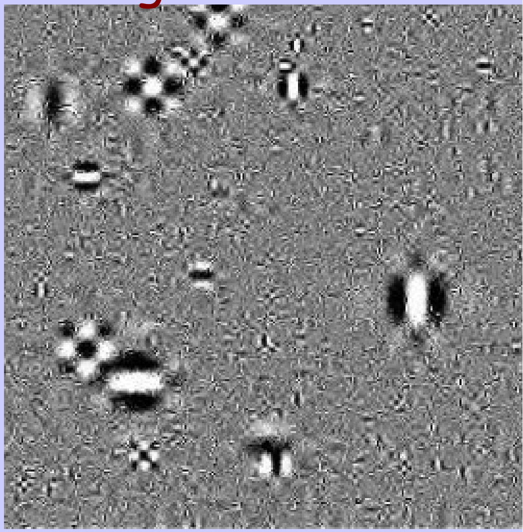


Sampling High-Resolution Images

Gaussian model

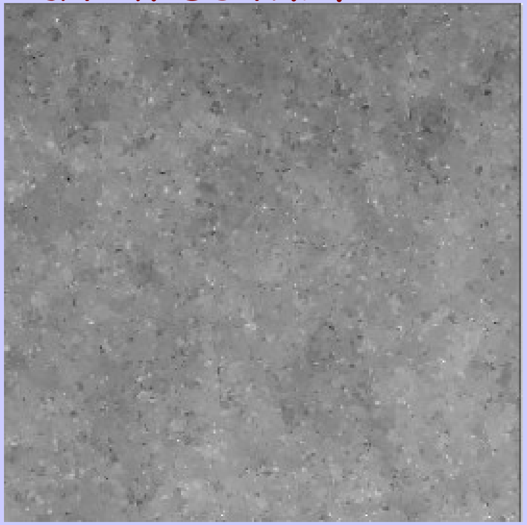


marginal wavelet



from Simoncelli 2005

Pair-wise MRF



FoE

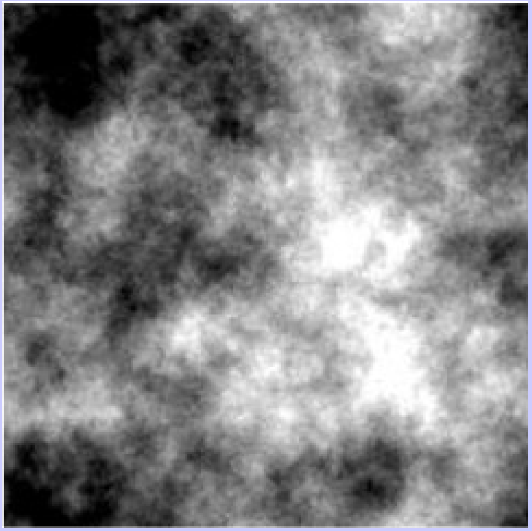
from Schmidt, Gao, Roth CVPR 2010

Sampling High-Resolution Images

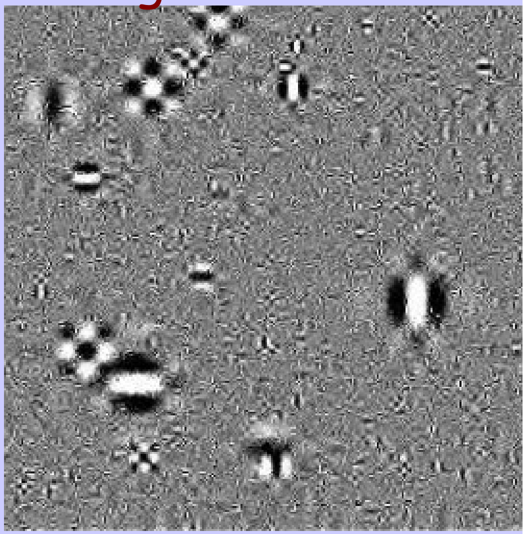
Mean Covariance Model

Ranzato, Mnih, Hinton NIPS 2010

Gaussian model

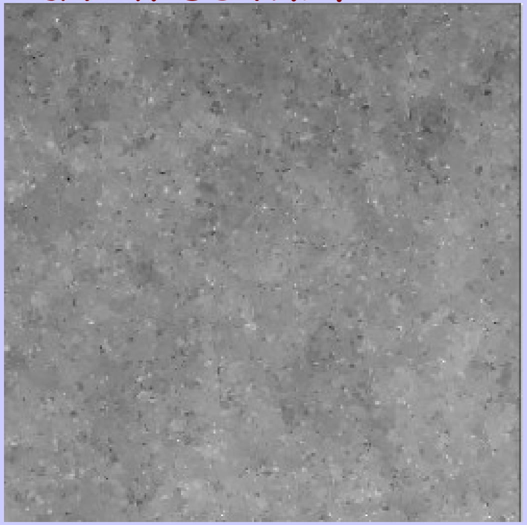


marginal wavelet



from Simoncelli 2005

Pair-wise MRF



FoE

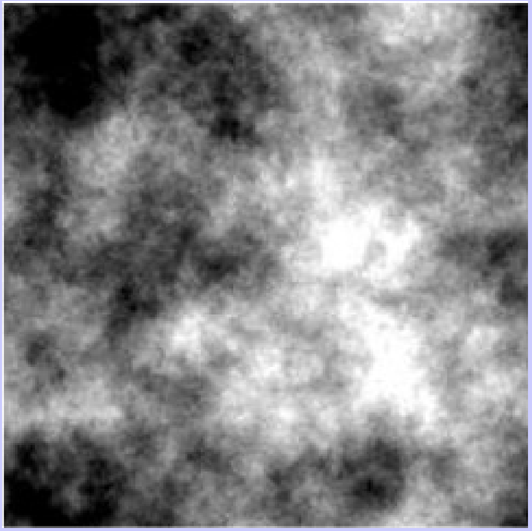
from Schmidt, Gao, Roth CVPR 2010

Sampling High-Resolution Images

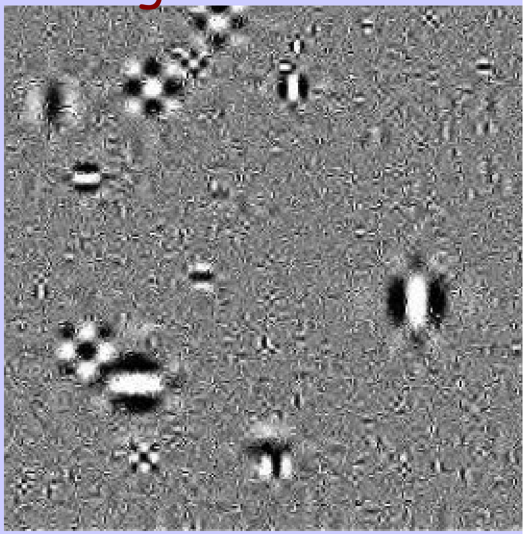
Mean Covariance Model

Ranzato, Mnih, Hinton NIPS 2010

Gaussian model

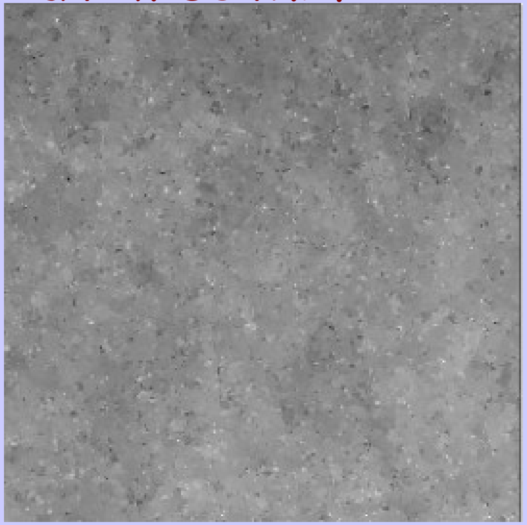


marginal wavelet



from Simoncelli 2005

Pair-wise MRF

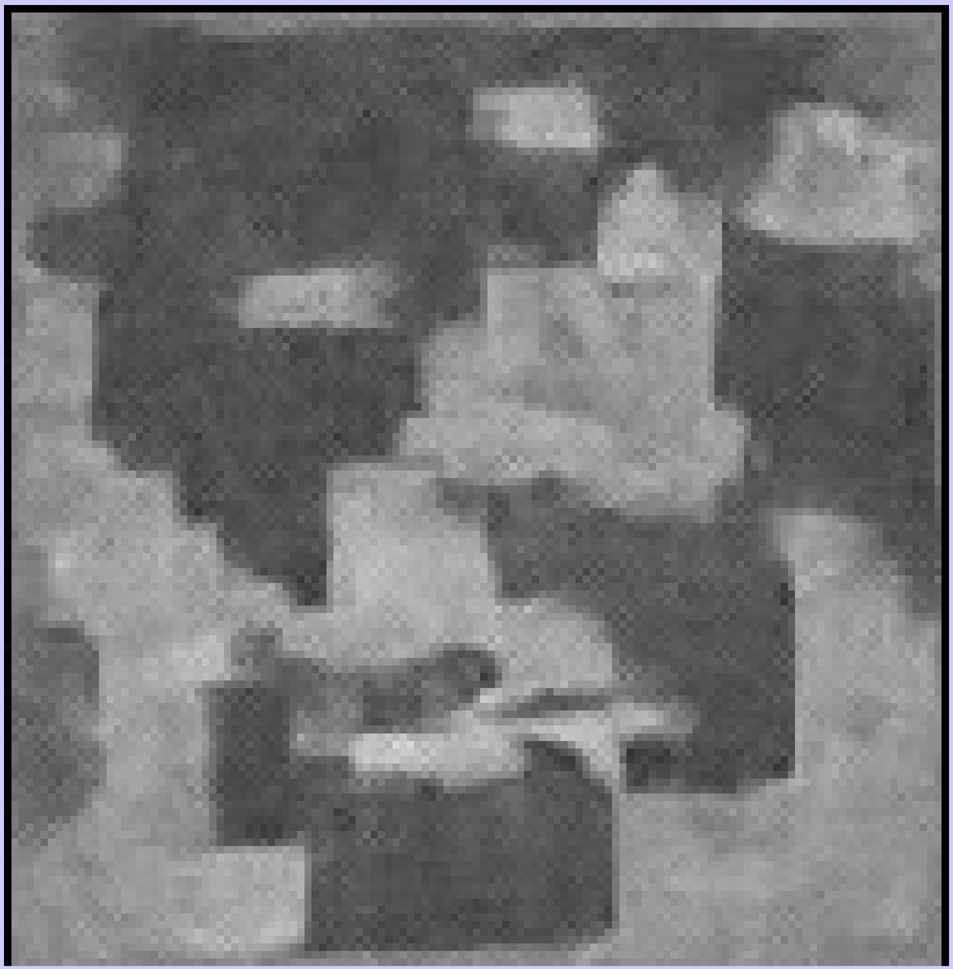


FoE

from Schmidt, Gao, Roth CVPR 2010

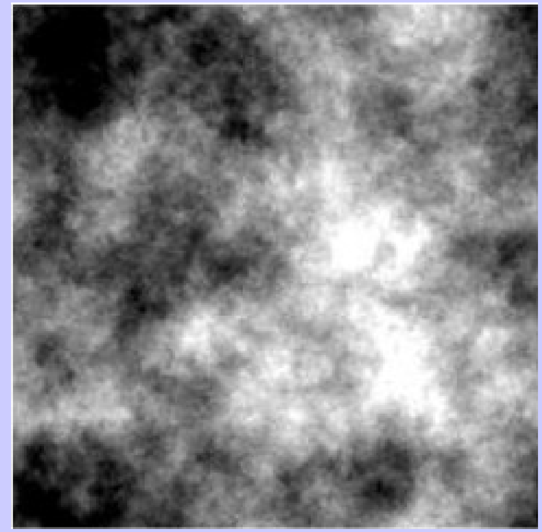
Sampling High-Resolution Images

Mean Covariance Model

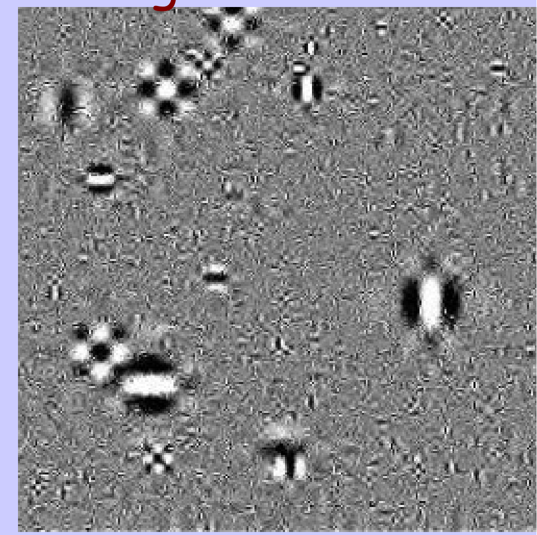


Ranzato, Mnih, Hinton NIPS 2010

Gaussian model

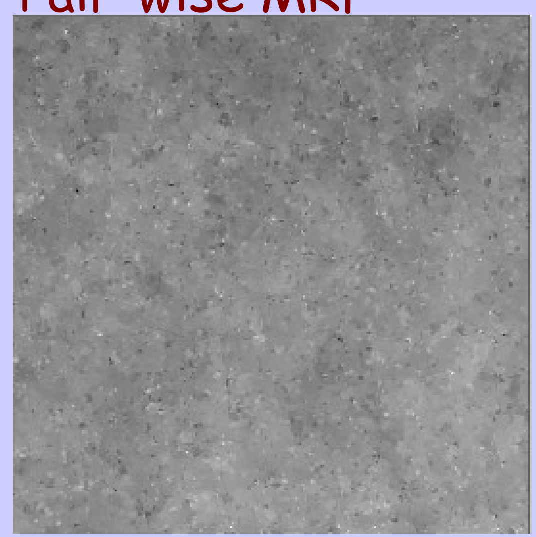


marginal wavelet

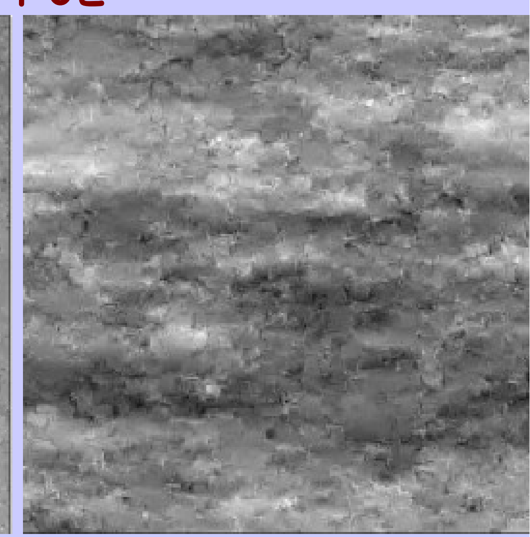


from Simoncelli 2005

Pair-wise MRF



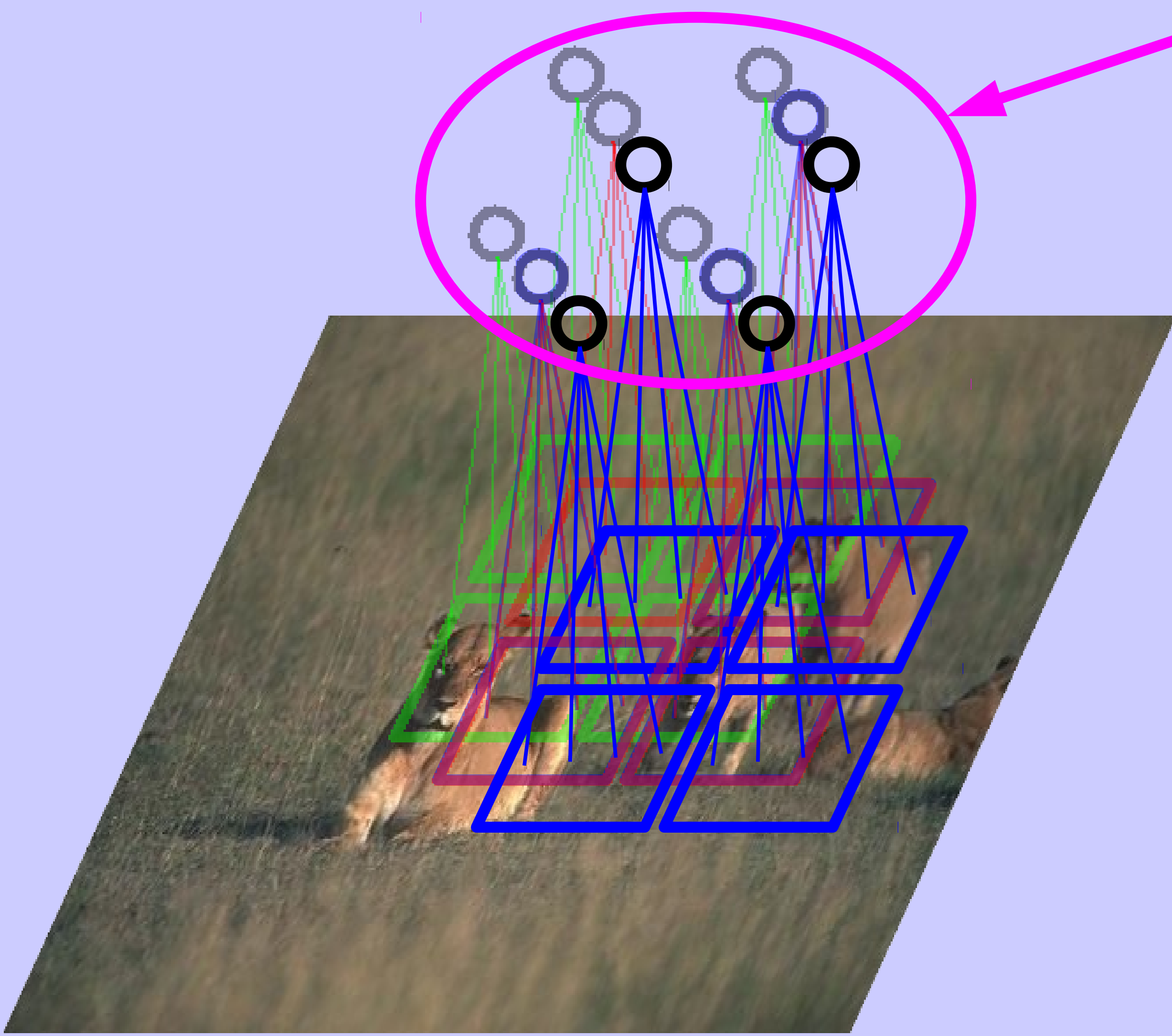
FoE



from Schmidt, Gao, Roth CVPR 2010

Making the model.. "DEEPER "

Treat these units as data
to train a similar model on the top

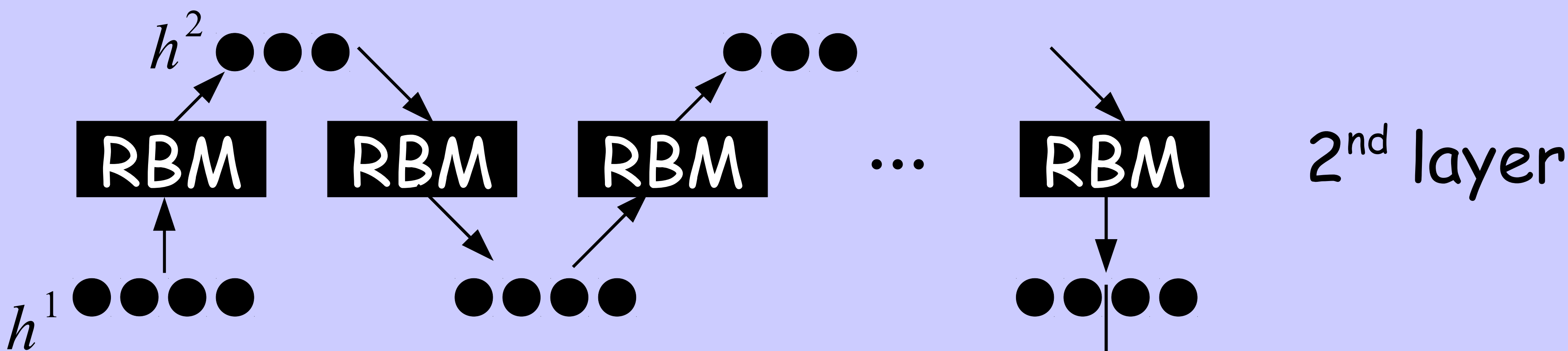


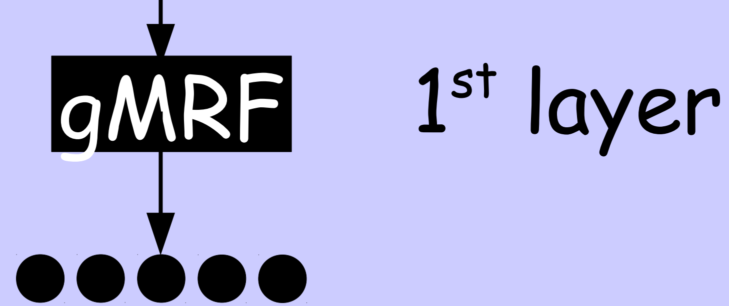
SECOND STAGE

Field of binary RBM's.
Each hidden unit has a
receptive field of 30x30
pixels in input space.

Sampling from the DEEPER model

- Sample from 2nd layer Restricted Boltzmann Machine (RBM)
- project sample in image space using 1st layer p(x|h)



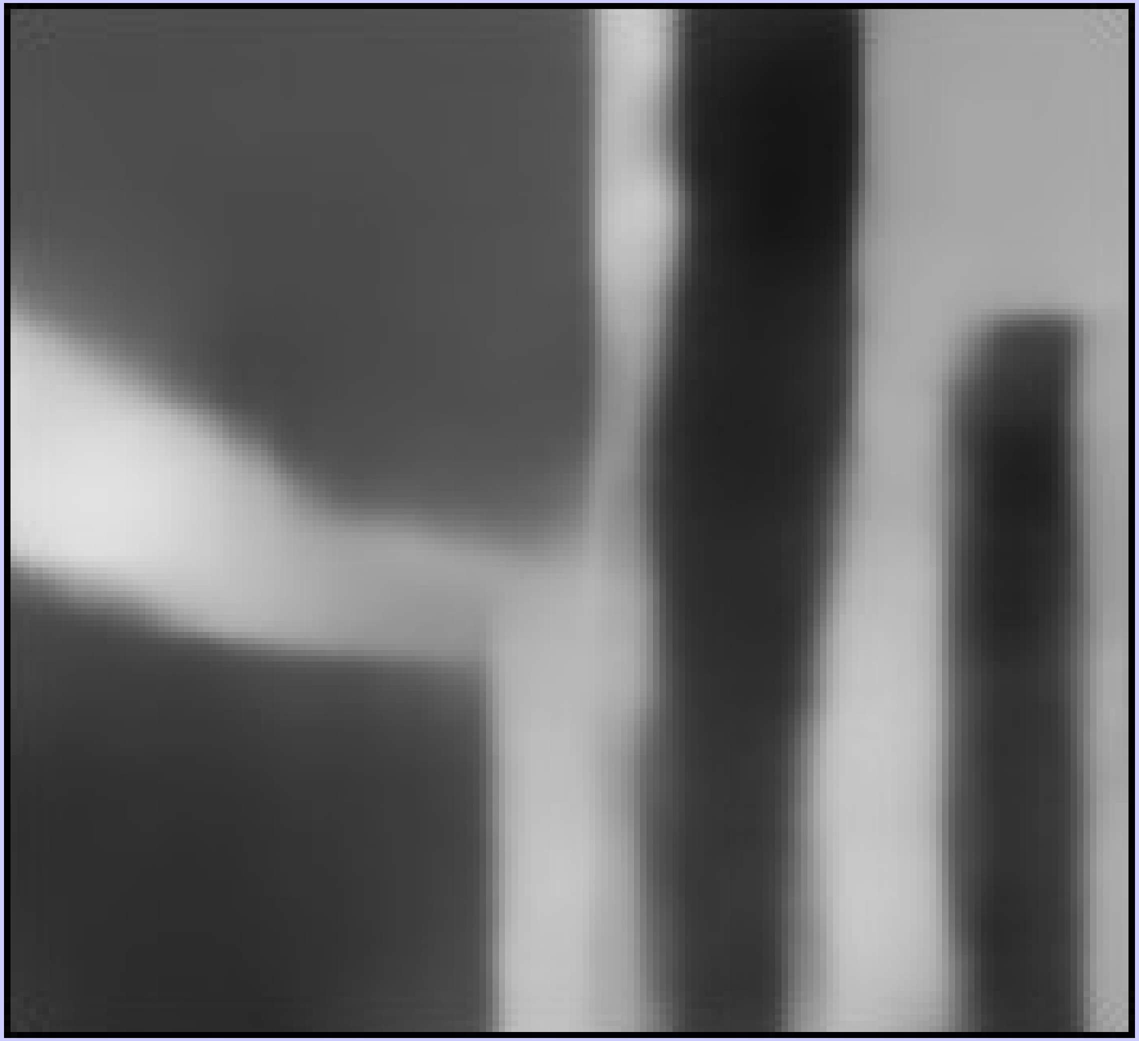


$$E(h^1, h^2) = -h^{1\prime} W h^2$$
$$p(h_j^2 = 1 | h^1) = \sigma(W_j{}' h^1 + b_j^2)$$
$$p(h_k^1 = 1 | h^2) = \sigma(W_k h^2 + b_k^1)$$

Samples from Deep Generative Model

1st stage model

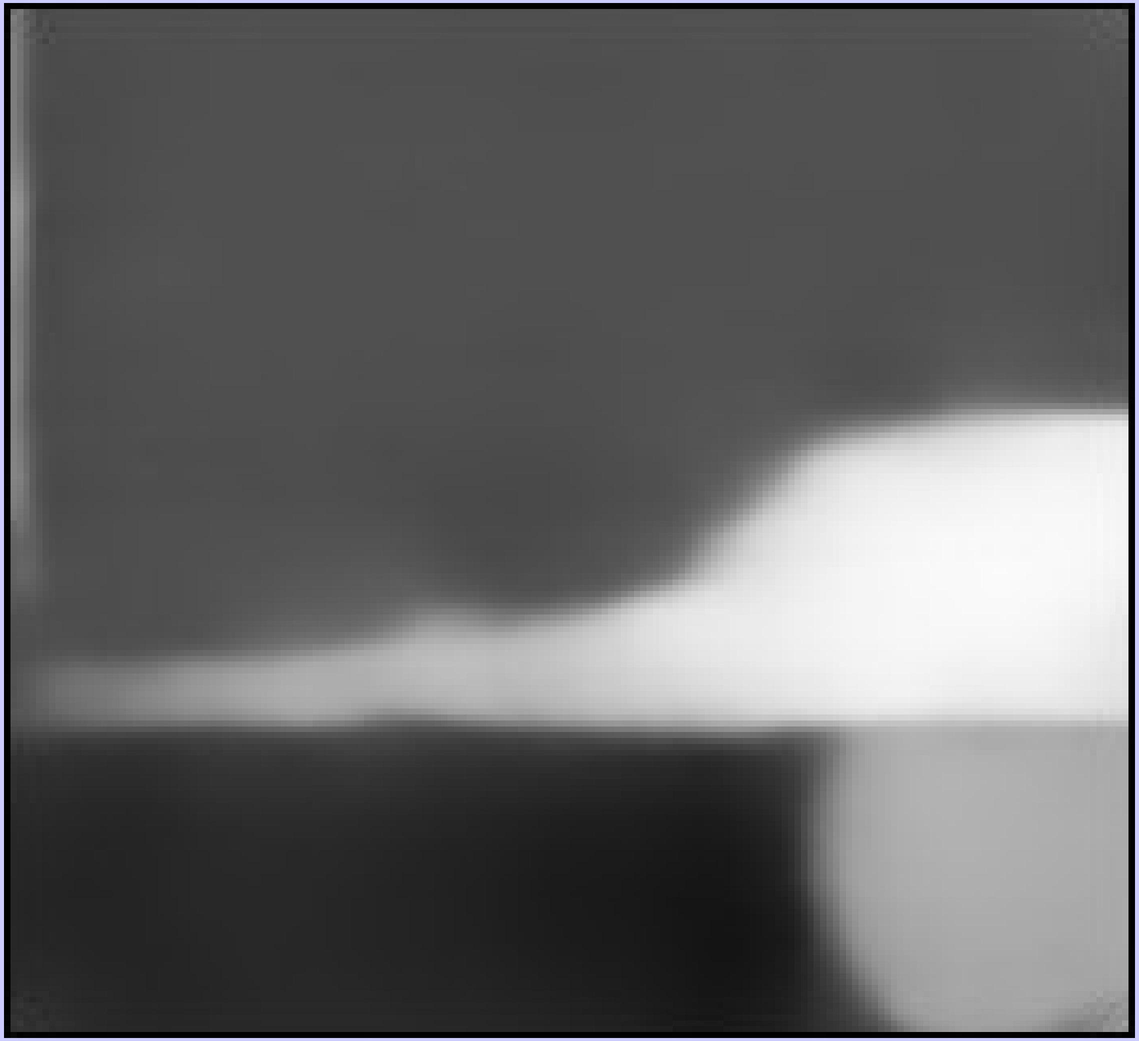
3rd stage model



Samples from Deep Generative Model

1st stage model

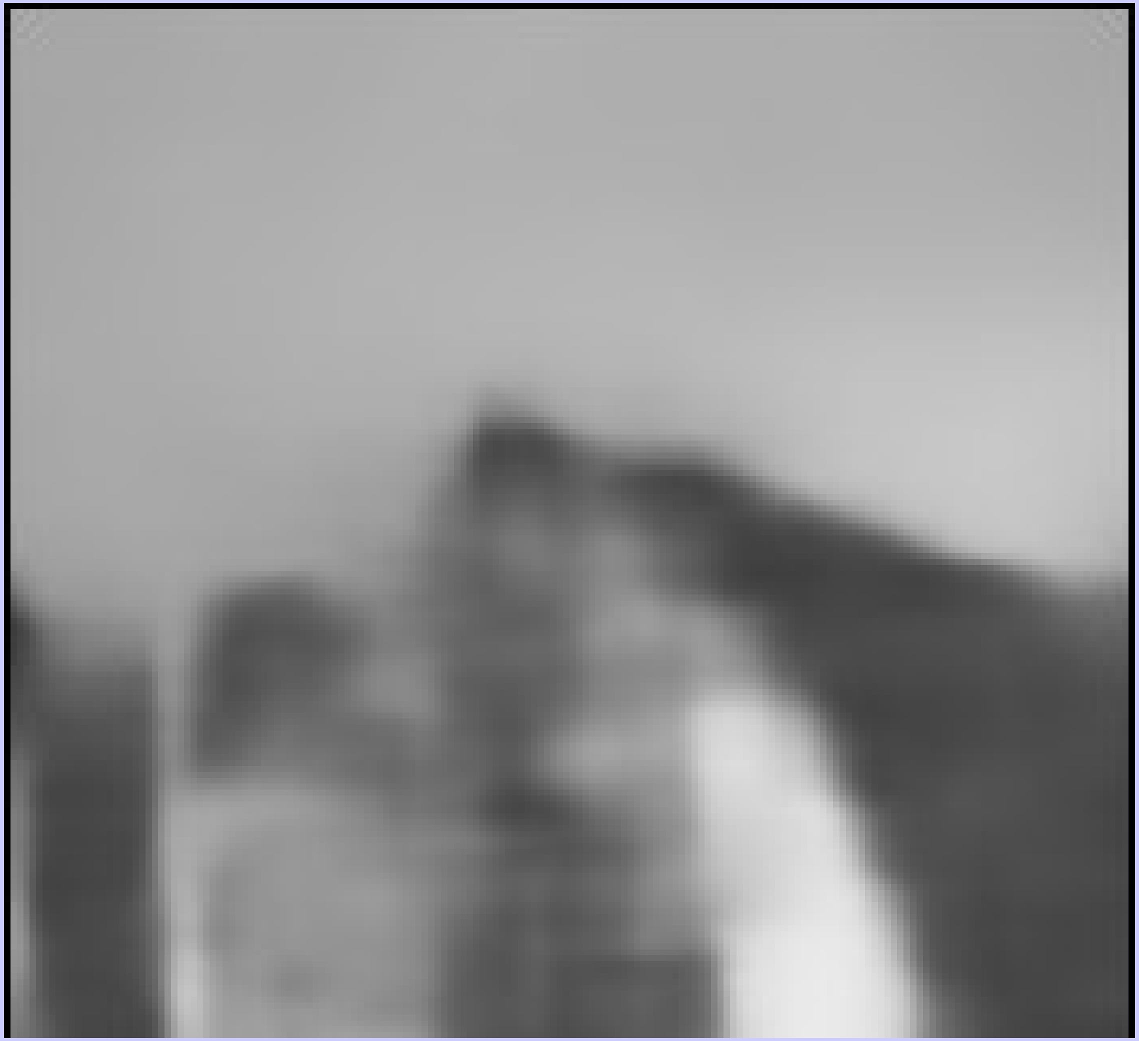
3rd stage model



Samples from Deep Generative Model

1st stage model

3rd stage model



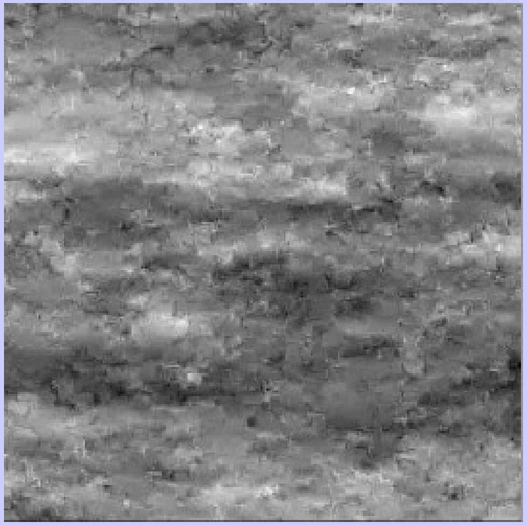
Samples from Deep Generative Model

1st stage model

3rd stage model

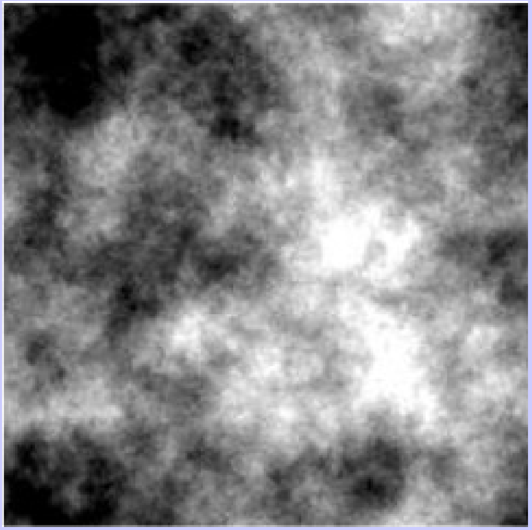
Sampling High-Resolution Images

FoE



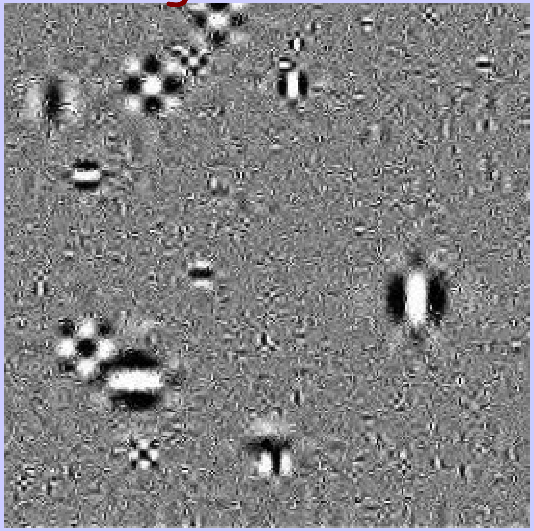
from Schmidt, Gao, Roth CVPR 2010

Gaussian model



from Simoncelli 2005

marginal wavelet

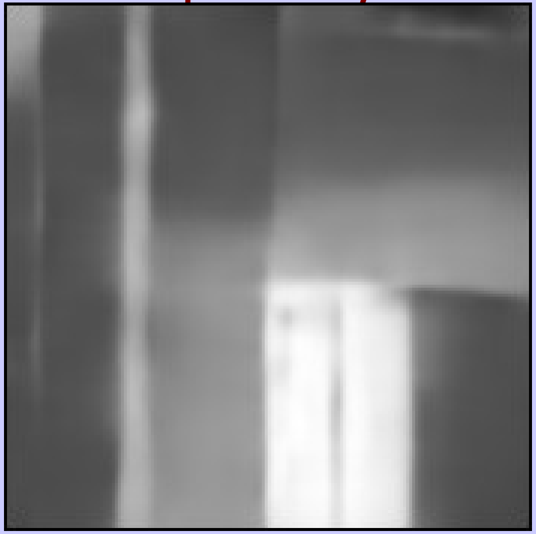


from Simoncelli 2005

Deep - 1 layer

Ranzato, Mnih, Hinton NIPS 2010

Deep - 3 layers



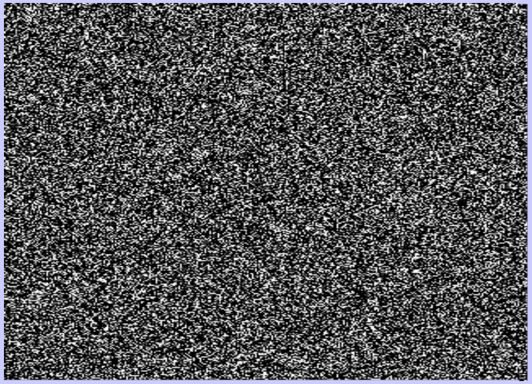
Ranzato, et al. CVPR 2011

Using -Energy to Score Images

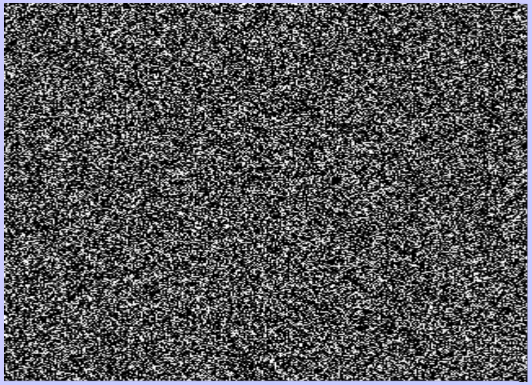
less likely

test images

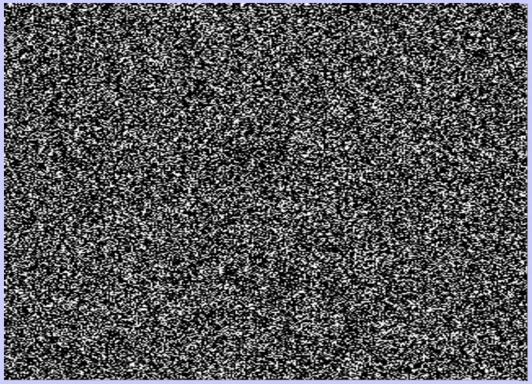
>



>

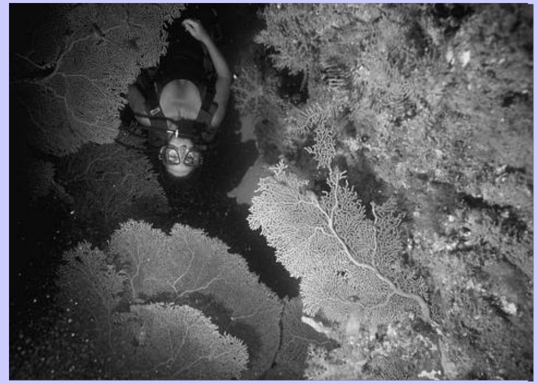
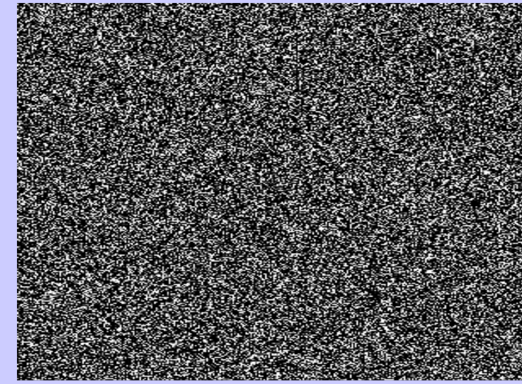


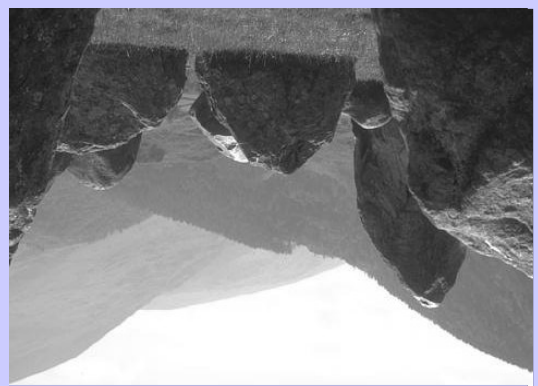
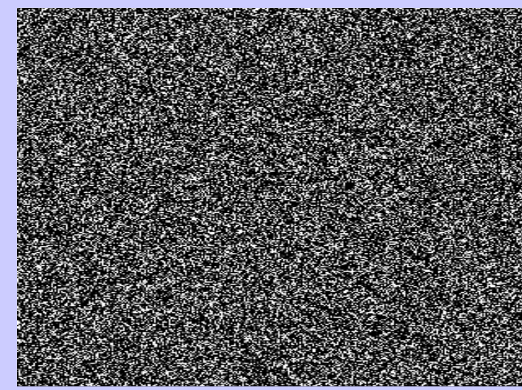
>

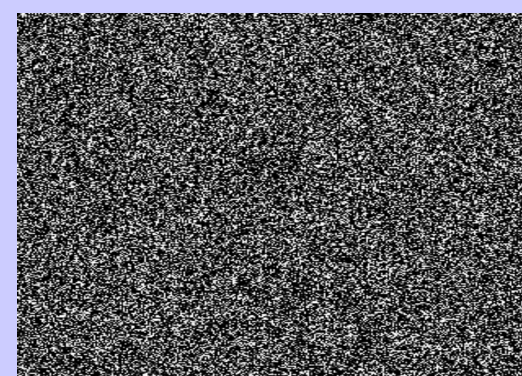


Using Energy to Score Images

Upside-down images

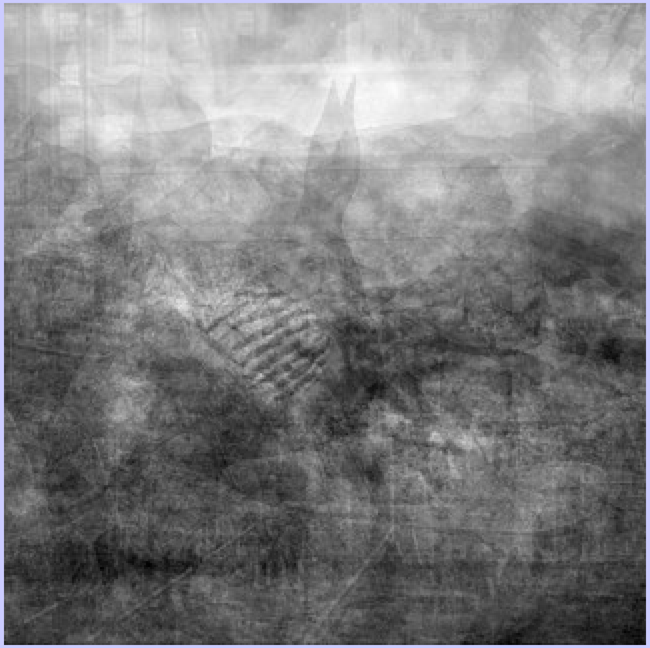
 >  > 

 >  > 

 >  > 

Using Energy to Score Images

Average of those images for which difference of energy is higher



Scene Recognition

- 15 scene dataset (Lazebnik et al. CVPR 2006)
  - 15 categories, 100 images per class for training

Scene Recognition

- use hiddens at 2nd layer to represent 46x46 input image patches
- spatial pyramid matching on 1st and 2nd layer fearures

- Result

accuracy non-linear SVM (histogram intersection)

- SIFT...81.4%

Lazebnik et al. CVPR 2006

- DEEP Features: 81.2%

Ranzato et al. CVPR 2011

- Best Method (SIFT + Sparse Coding)...............................84.1%

Boureau et al. CVPR 2010

Image Denoising

original image

noisy image: PSNR=22.1dB

denoised: PSNR=28.0dB

$$X^* = \operatorname{argmin} \frac{1}{2} \frac{\|X - N\|}{\sigma^2} + F(X)$$

Image Denoising

original image

noisy image: PSNR=22.1dB

denoised: PSNR=29.2dB

repeat

$$X^* = argmin \frac{1}{2} \frac{\|X - N\|}{\sigma^2} + F(X)$$

$$\theta^* = argmin_\theta - \log p(X^*; \theta)$$

Image Denoising

original image

noisy image: PSNR=22.1dB

denoised: PSNR=30.7dB

$$X^* = \alpha X^*_{mPoT} + (1-\alpha) X^*_{NonLocalMeans}$$

Outline

- mathematical formulation of the model
- training
- generation of natural images
- recognition of facial expression under occlusion
- learning acoustic features for spech recognition
- conclusion

Facial Expression Recognition

Toronto Face Dataset (J. Susskind et al. 2010)

~ 100K unlabeled faces from different sources

~ 4K labeled images

Resolution: 48x48 pixels

7 facial expressions

anger

Ranzato, et al. CVPR 2011

Facial Expression Recognition

Toronto Face Dataset (J. Susskind et al. 2010)

~ 100K unlabeled faces from different sources

~ 4K labeled images

Resolution: 48x48 pixels

7 facial expressions

disgust

Facial Expression Recognition

Toronto Face Dataset (J. Susskind et al. 2010)

~ 100K unlabeled faces from different sources

~ 4K labeled images

Resolution: 48x48 pixels

7 facial expressions

fear

Facial Expression Recognition

Toronto Face Dataset (J. Susskind et al. 2010)

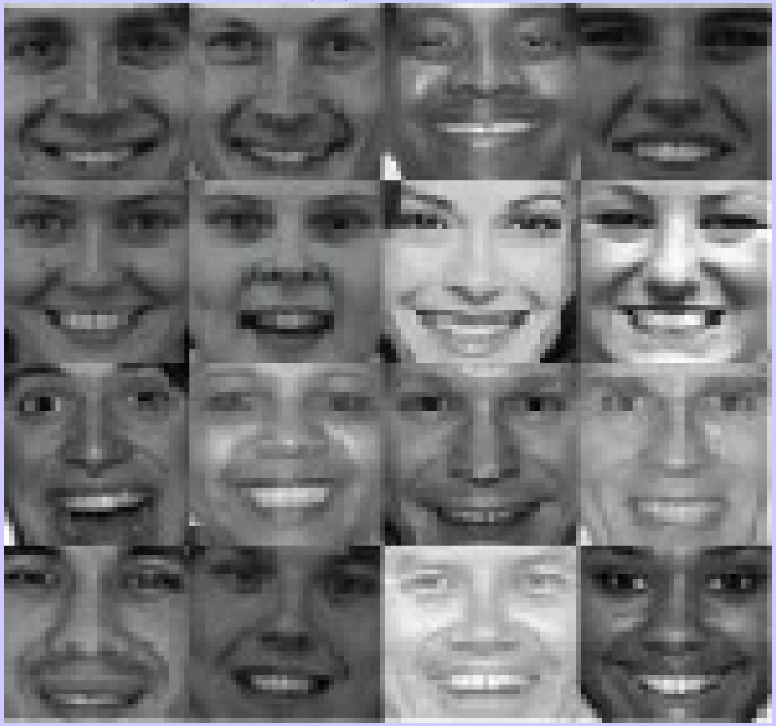
~ 100K unlabeled faces from different sources

~ 4K labeled images

Resolution: 48x48 pixels

7 facial expressions

happiness



Facial Expression Recognition

Toronto Face Dataset (J. Susskind et al. 2010)

~ 100K unlabeled faces from different sources

~ 4K labeled images

Resolution: 48x48 pixels

7 facial expressions

neutral

Facial Expression Recognition

Toronto Face Dataset (J. Susskind et al. 2010)

~ 100K unlabeled faces from different sources

~ 4K labeled images

Resolution: 48x48 pixels

7 facial expressions

sadness

Facial Expression Recognition

Toronto Face Dataset (J. Susskind et al. 2010)

~ 100K unlabeled faces from different sources

~ 4K labeled images

Resolution: 48x48 pixels

7 facial expressions

surprise

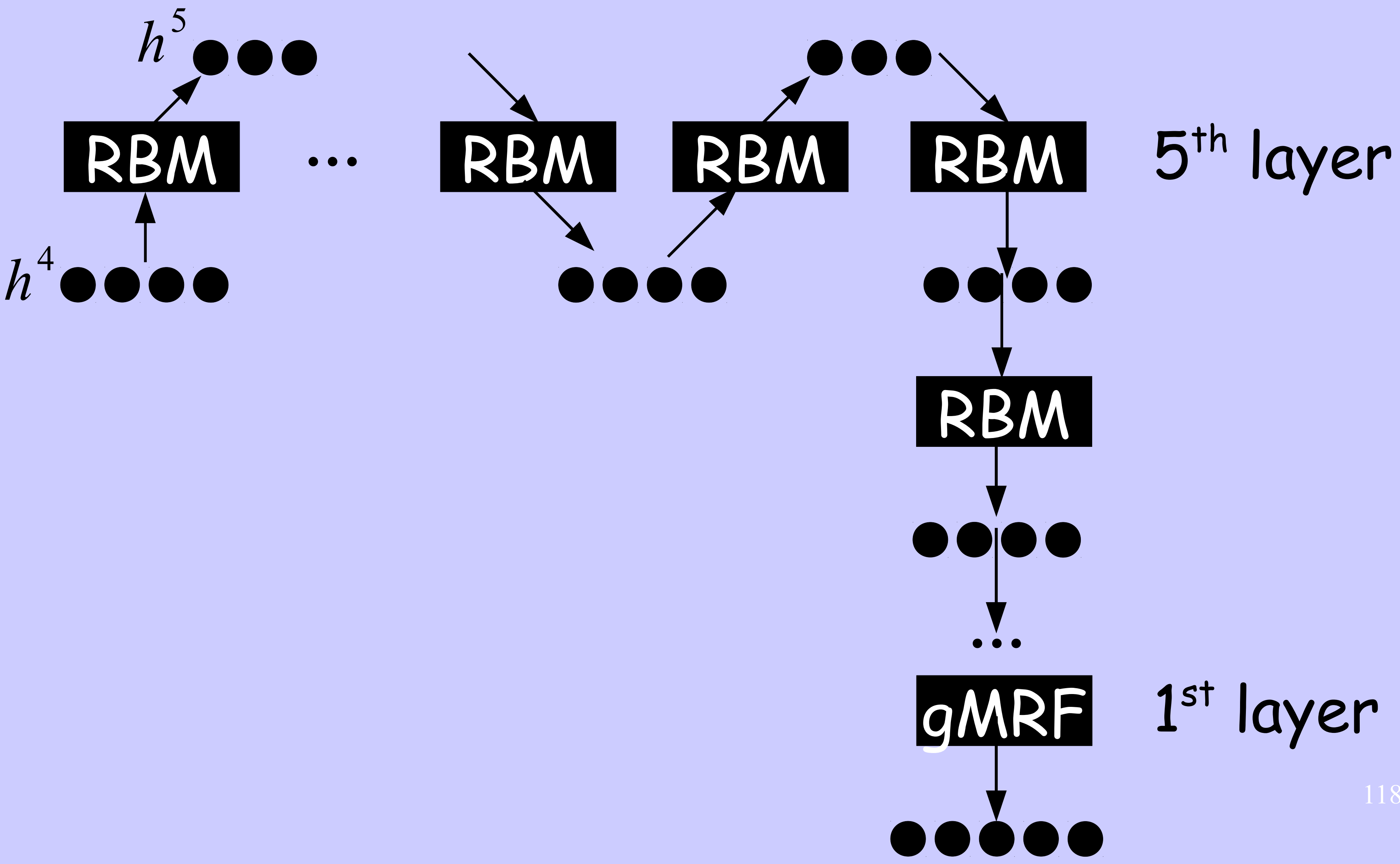
Facial Expression Recognition

- 1st layer using local (not shared) connectivity
- layers above are fully connected
- 5 layers in total

- Result
  - Linear Classifier on raw pixels 71.5%
  - Gaussian RBF SVM on raw pixels 76.2%
  - Gabor + PCA + linear classifier 80.1%
    Dailey et al. J. Cog. Science 2002
  - Sparse coding 74.6%
    Wright et al. PAMI 2008
  - DEEP model (3 layers): 82.5%

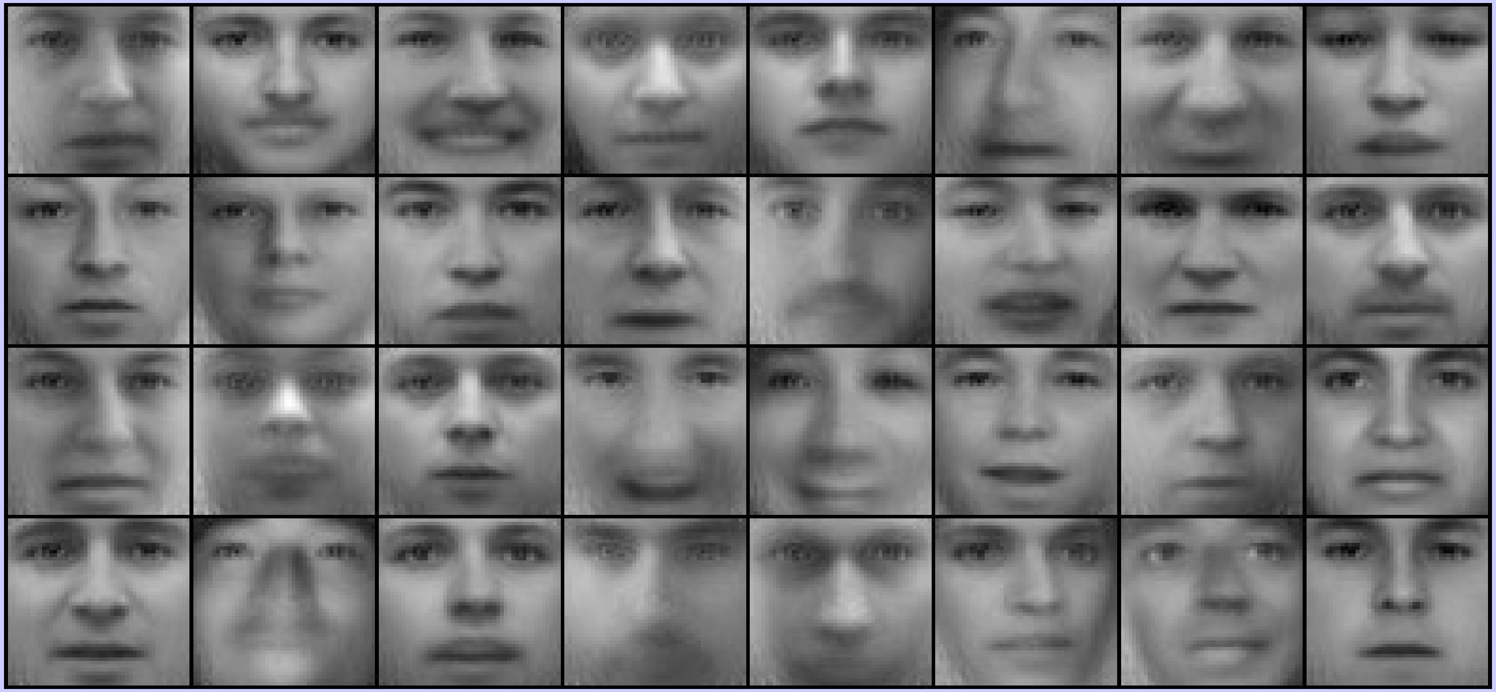
Facial Expression Recognition

Drawing samples from the model (5th layer with 128 hiddens)



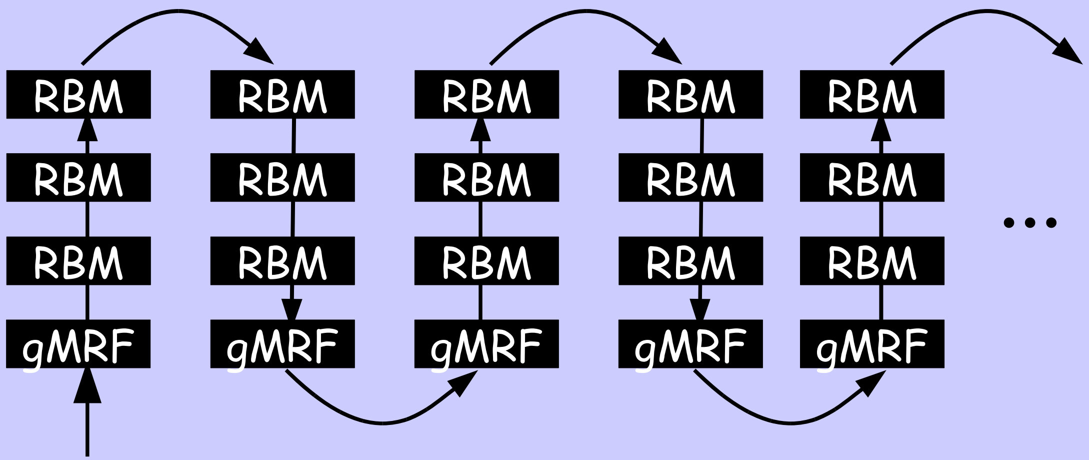
Facial Expression Recognition

Drawing samples from the model (5^{th} layer with 128 hiddens)



Facial Expression Recognition

- 7 synthetic occlusions
- use generative model to fill-in
  (conditional on the known pixels)



Facial Expression Recognition

originals

Type 1 occlusion: eyes

Restored images

Facial Expression Recognition

originals

Type 2 occlusion: mouth

Restored images

Facial Expression Recognition

originals

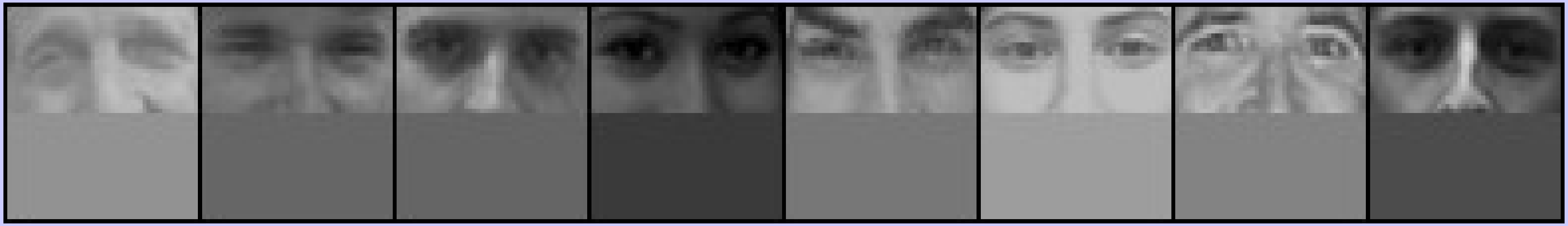
Type 3 occlusion: right half

Restored images

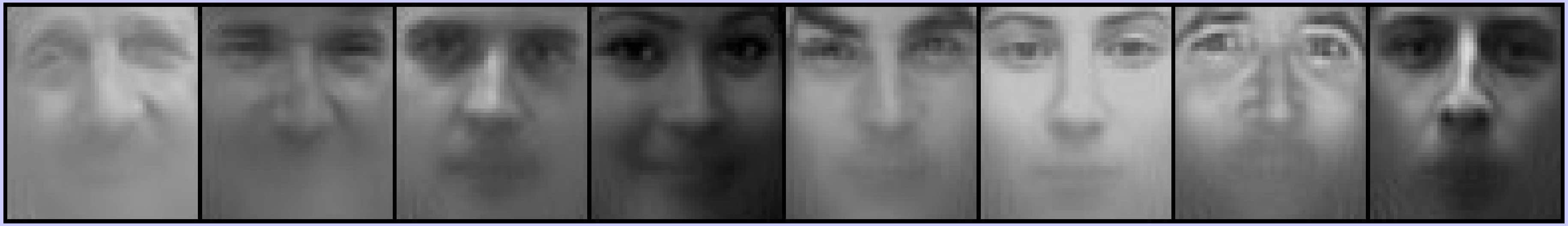
Facial Expression Recognition

originals

Type 4 occlusion: bottom half



Restored images

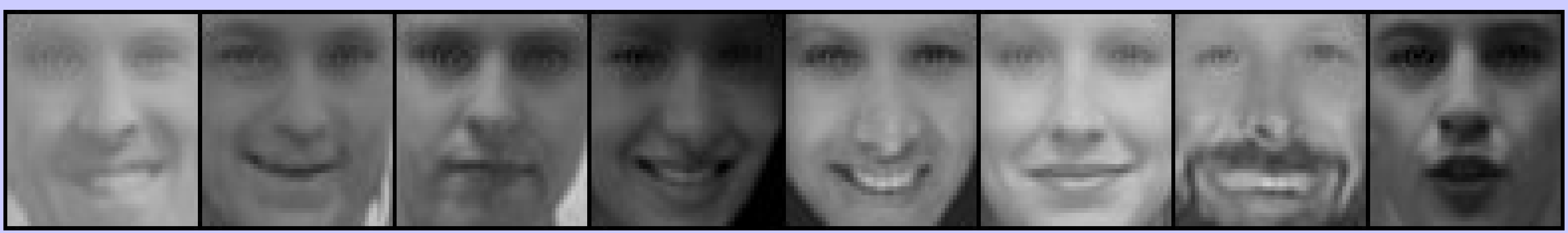


Facial Expression Recognition

originals

Type 5 occlusion: top half

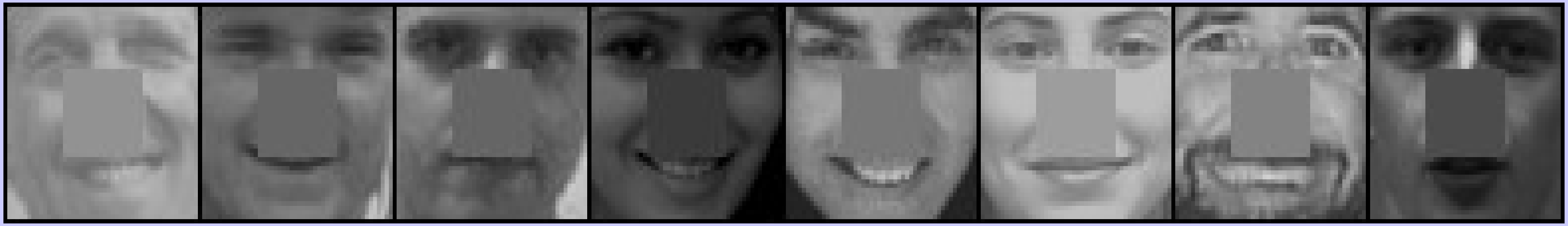
Restored images



Facial Expression Recognition

originals

Type 6 occlusion: nose

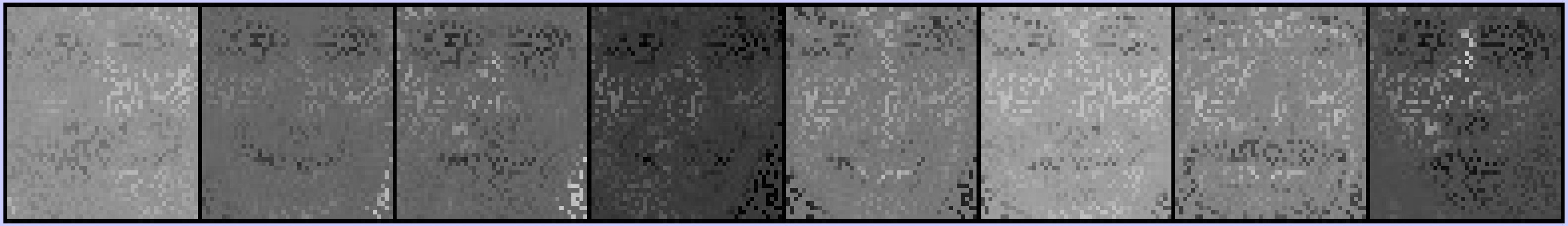


Restored images

Facial Expression Recognition

originals

Type 7 occlusion: 70% of pixels at random



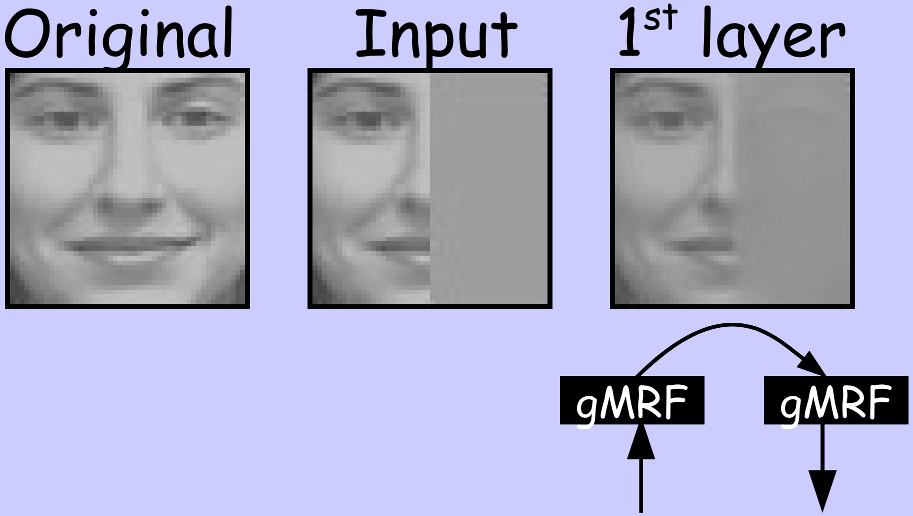
Restored images

Facial Expression Recognition

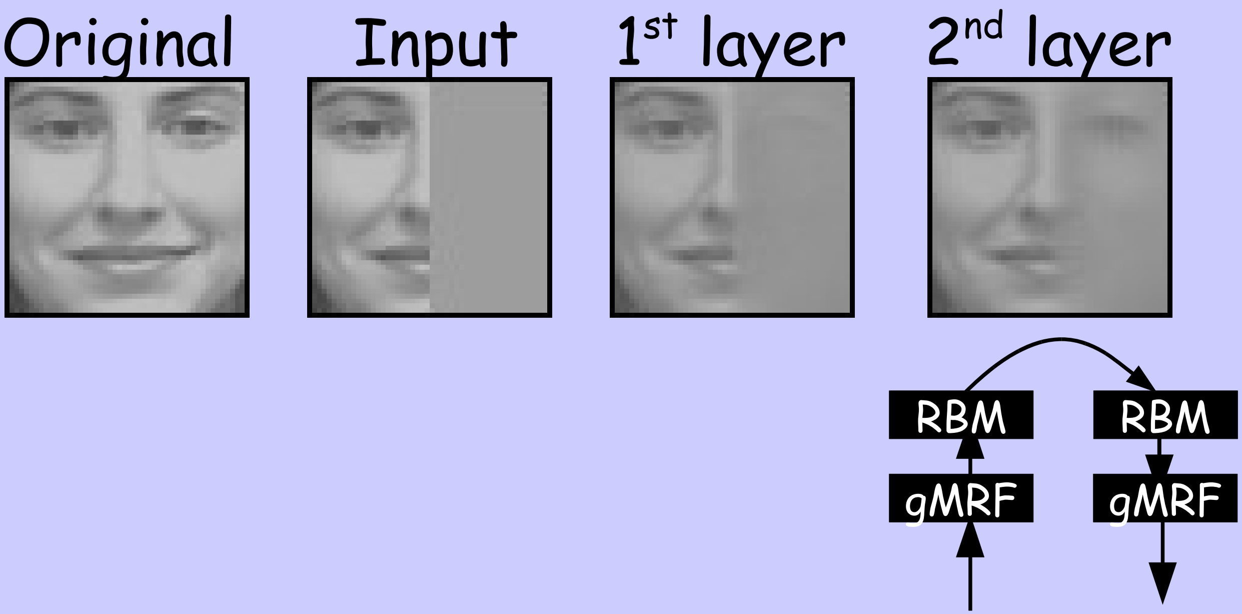
Original

Input

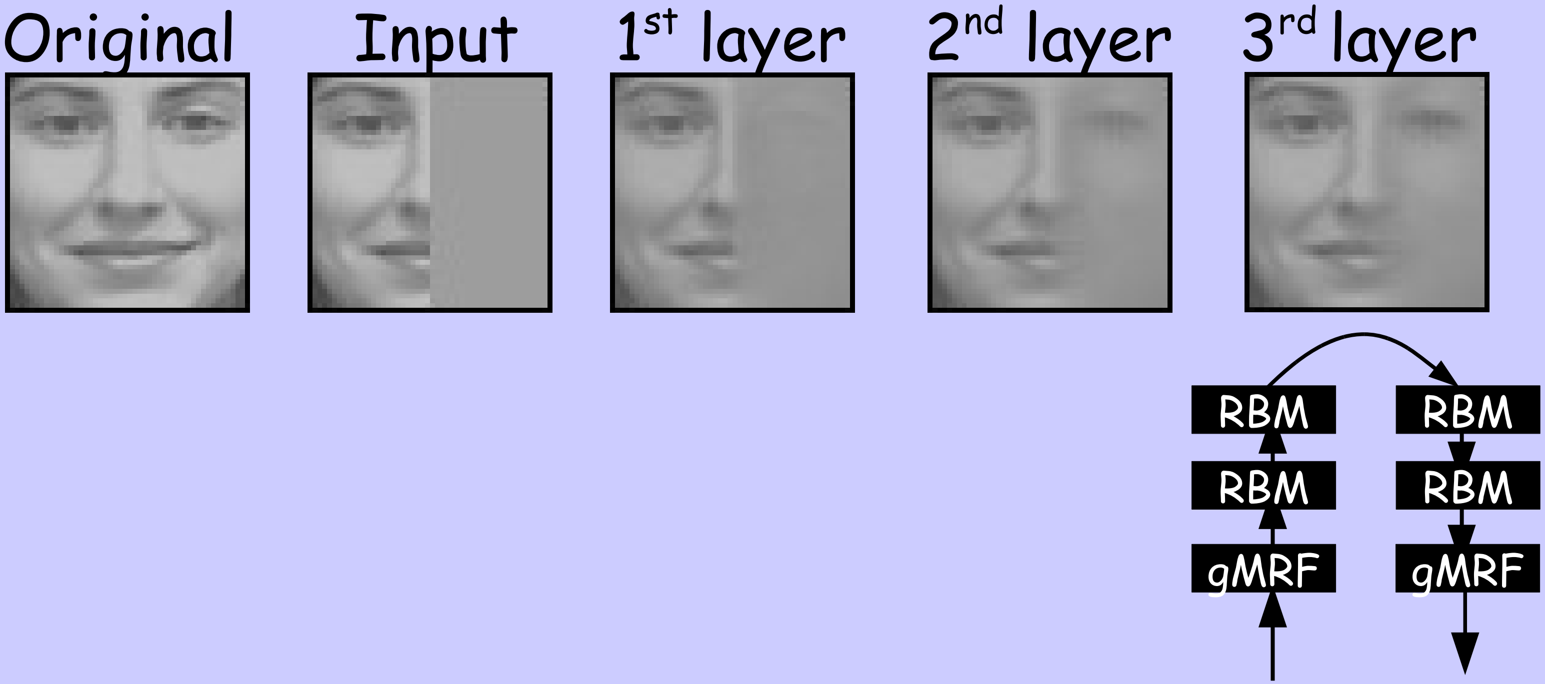
Facial Expression Recognition



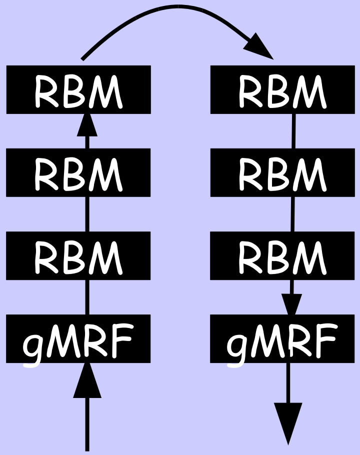
Facial Expression Recognition



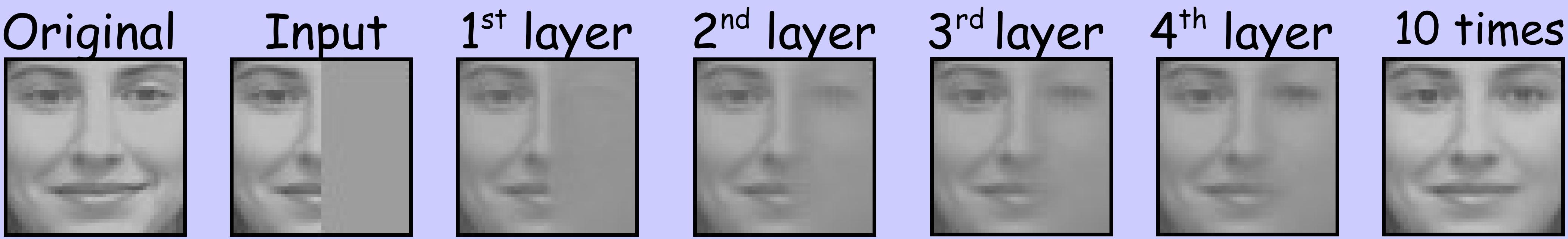
Facial Expression Recognition

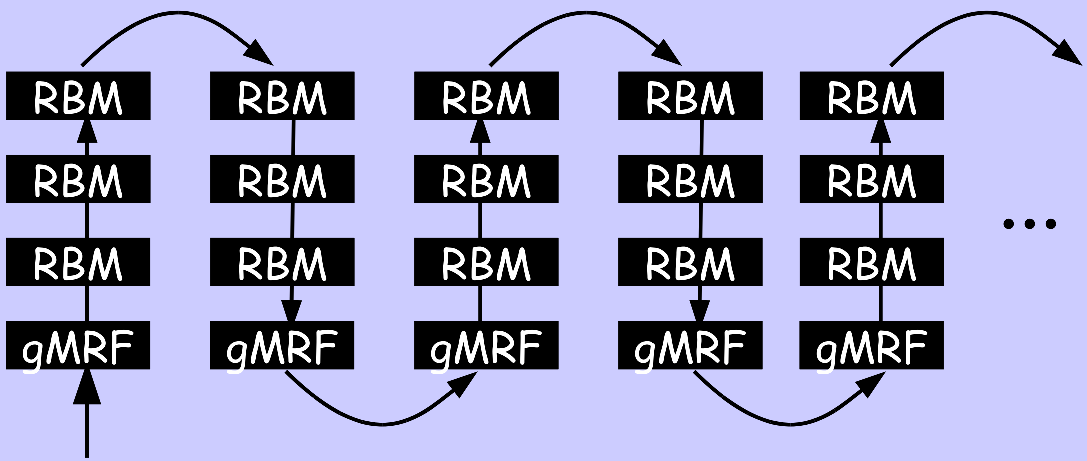


Facial Expression Recognition



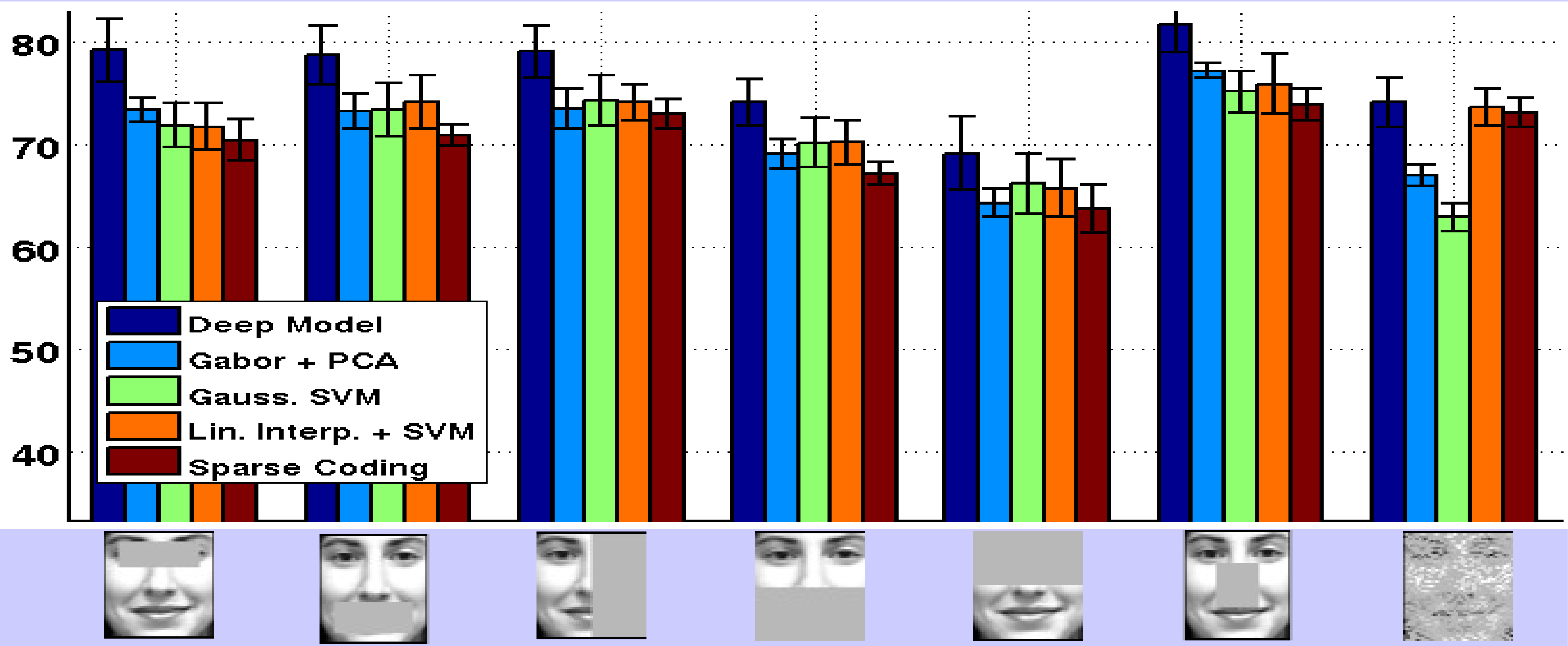
Facial Expression Recognition





Facial Expression Recognition

occluded images for both training and test



Dailey, et al. J. Cog. Neuros. 2003

Wright, et al. PAMI 2008

Ranzato, et al. CVPR 2011

Outline

- mathematical formulation of the model

- training

- generation of natural images

- recognition of facial expression under occlusion

- learning acoustic features for spech recognition

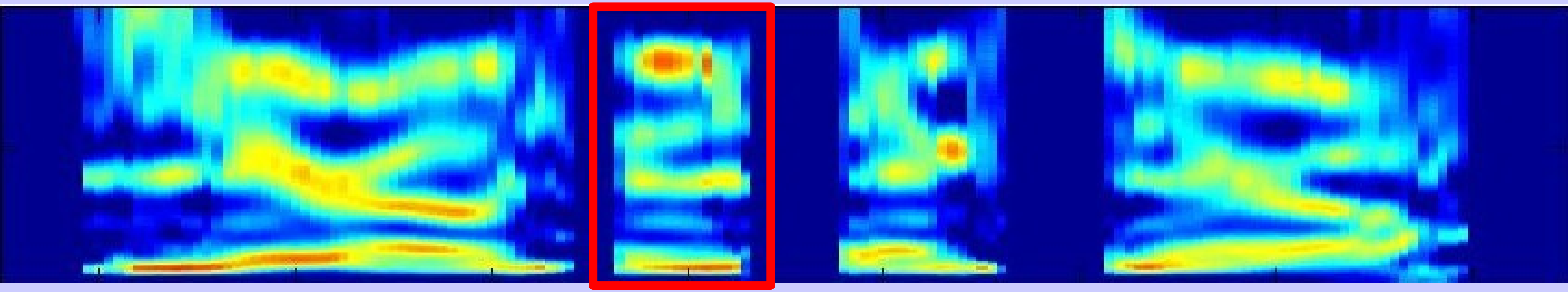
- conclusion

Speech Recognition on TIMIT

INPUT: standard pre-processing, but without augmentation
(no 1st & 2nd order termporal derivatives)

Training:
- unsupervised layer-wise training (8 layers, ~2000 units per layer)
- supervised training to predict states of HMM

Test: frame-by-frame prediction → Viterbi decoding



<-235 ms->

Dahl, Ranzato, Mohamed, Hinton, NIPS 2010

Speech Recognition on TIMIT

METHOD	PER
CRF	34.8%
Large-Margin GMM	33.0%
CD-HMM	27.3%
Augmented CRF	26.6%
RNN	26.1%
Bayesian Triphone HMM	25.6%
Triphone HMM discrim. trained	22.7%
DBN with gated MRF	20.5%

Speech Recognition on TIMIT

METHOD	PER	Year
CRF	34.8%	*2008*
Large-Margin GMM	33.0%	*2006*
CD-HMM	27.3%	*2009*
Augmented CRF	26.6%	*2009*
RNN	26.1%	*1994*
Bayesian Triphone HMM	25.6%	*1998*
Triphone HMM discrim. trained	22.7%	*2009*
DBN with gated MRF	20.5%	*2010*

Summary

- Unsupervised Learning
- Deep Generative Model
  - 1^{st} layer: gated MRF
  - Higher layers: binary RBM's
  - fast inference
  - Realistic generation: natural images
  - Applications:
    - scene recognition, denoising, facial expression recognition robust to occlusion...
    - speech recognition

THANK YOU

References on gated MRFs

- PoT like models for modeling natural images
  - Hinton, the - Discovering multiple constraints that are frequently approximately satisfied UAI 2001
  - Welling, Hinton, Osindero - Learning sparse topographic representations with products of student's t distributions NIPS 2003
  - Teh, Welling, Osindero, Hinton – Energy-based models for sparse overcomplete representations JMLR 2003
  - Osindero, Welling, Hinton – Topographic product models applied to natural scene statistics Neural Comp. 2006
  - Roth, Black – Field of Experts IJCV 2009
  - Ranzato, Krizhevsky, Hinton – Factored 3-way RBMs for modeling natural images AISTATS 2010
- mPoT like models for modeling images and speech
  - Ranzato, Hinton – Modeling pixel means and covariances using factored 3rd order Boltzmann machines CVPR 2010
  - Dahl, Ranzato, Mohamed, Hinton – Phone recognition with mcRBM NIPS 2010
  - Ranzato, Mnih, Hinton – Generating more realistic images using gated MRF's NIPS 2010
  - Ranzato, Susskind, Mnih, Hinton – On deep generative models with applications to recognition CVPR 2011
  - Kivinen, Williams – Multiple texture Boltzmann machines AISTATS 2012

- Models similar to mPoT
  - Courville, Bergstra, Bengio – The spike and slab RBM NIPS 2010
  - Courville, Bergstra, Bengio – Unsupervised models of image by ssRBM ICML2011
  - Goodfellow, Courville, Bengio – Large-scale feature learning with spike-and-slab sparse coding. ICML 2012
- 3-way RBM applied to sequences
  - Memisevic, Hinton – Unsupervised learning of image transformations CVPR 2007
  - Taylor, Hinton – Factored conditional RBM for modeling motion style ICML 2009
  - Memisevic, Hinton – Learning to represent spatial transformations with a factored high-order Boltzmann machine Neural Comp 2010
  - Memisevic – Gradient-based learning of higher-order image features ICCV 2011
  - Memisevic – On multi-view feature learning ICML 2012