



# Data modeling with Energy Based Models

**Beatriz Seoane**

*LISN Paris-Saclay University*

# Acknowledgments

Aurélien Decelle  
Giovanni Catania  
Alfonso Navas  
Lorenzo Rosset



Université  
PARIS-SACLAY

Nicolas Béreux  
Cyril Furtlechner



MINISTERIO  
DE CIENCIA  
E INNOVACIÓN

el  
AGENCIA  
ESTATAL DE  
INVESTIGACIÓN

Santander

ANR  
AGENCE NATIONALE  
DE LA  
RECHERCHE

# Plan for the lecturers

- Class 1: **Introduction** to Energy Based Models
- Class 2: **Interpretability**. How can we learn from trained networks?
- Class 3: **Training optimization, the role of MCMC**. How can we improve the training mechanisms by understanding their physics?

# Plan for the lecturers

- Class 1: **Introduction** to Energy Based Models
  - Generative approach
  - Introduction to Energy-Based Models
    - The Restricted Boltzmann Machine (RBM)
  - Maximum likelihood training
  - Generation
  - Why I think RBMs are a cool tool

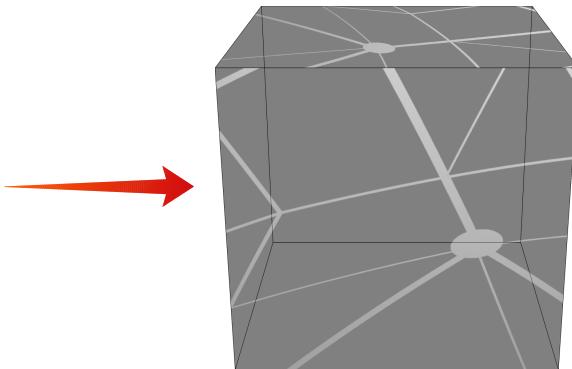


# General definitions

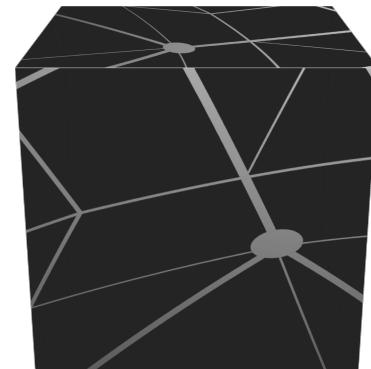
# Introduction : Generative approach

0 1 2 3 4 5 6 7 8 9

training



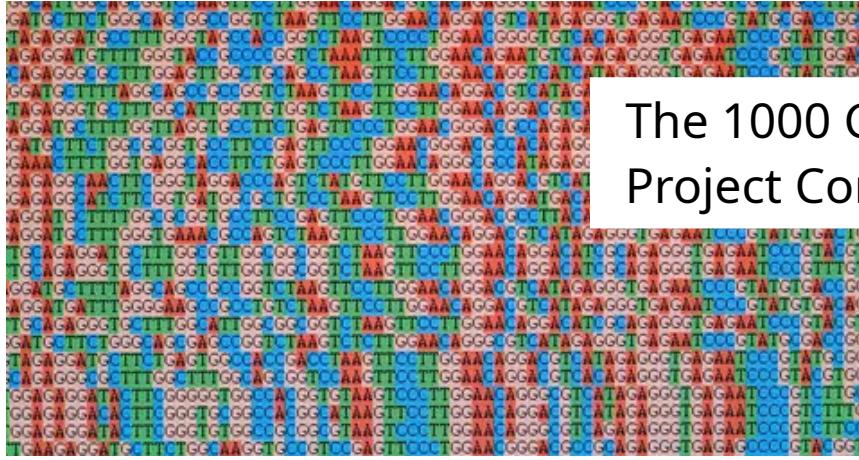
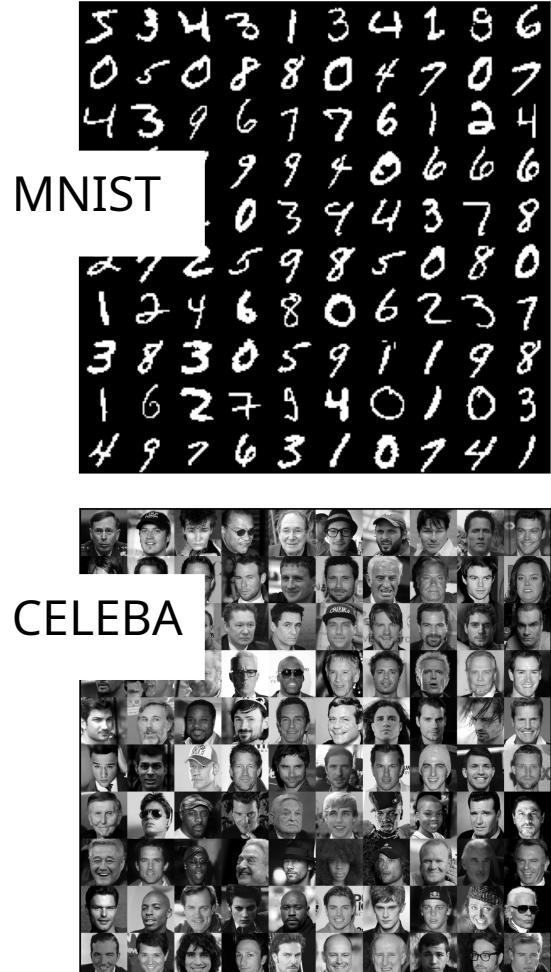
generating



7 6 5 4 3 2 1 0

- **Energy based models (RBMs, Generative Convnets)**
- **Diffusion models**, normalizing flows, score based
- Variational AutoEncoder (VAE)
- Generative Adverarial Network (GAN)
- Autoregressive methods

# Introduction : generative approach



Q5E940\_BOVIN ----- MPREDRATWPSNYFLKELIPLDDPKFCFVYGADNVGSKM0IIRMSLRKX-AVVLMGKRHNMRRKAIRGLHEENN--PALE 76  
RLAO\_HUMAN ----- MPREDRATWPSNYFLKELIPLDDPKFCFVYGADNVGSKM0IIRMSLRKX-AVVLMGKRHNMRRKAIRGLHEENN--PALE 76  
RLAO\_MOUSE ----- MPREDRATWPSNYFLKELIPLDDPKFCFVYGADNVGSKM0IIRMSLRKX-AVVLMGKRHNMRRKAIRGLHEENN--PALE 76  
RLAO\_RAT ----- MPREDRATWPSNYFLKELIPLDDPKFCFVYGADNVGSKM0IIRMSLRKX-AVVLMGKRHNMRRKAIRGLHEENN--PALE 76  
RLAO\_CHICK ----- MPREDRATWPSNYFMKELIPLDDPKFCFVYGADNVGSKM0IIRMSLRKX-AVVLMGKRHNMRRKAIRGLHEENN--PALE 76  
RLAO\_RAMSY ----- MPREDRATWPSNYFLKELIPLDDPKFCFVYGADNVGSKM0IIRMSLRKX-AVVLMGKRHNMRRKAIRGLHEENN--PALE 76  
Q7ZUG3\_BRARE ----- MPREDRATWPSNYFLKELIPLDDPKFCFVYGADNVGSKM0IIRMSLRKX-AVVLMGKRHNMRRKAIRGLHEENN--PALE 76  
RLAO\_ICTPU ----- MPREDRATWPSNYFLKELIPLDDPKFCFVYGADNVGSKM0IIRMSLRKX-AVVLMGKRHNMRRKAIRGLHEENN--PALE 76  
RLAO\_DROME ----- MYRENKRAAWFAQYFKVVEFDEEFKKCFVYGADNVGSKM0IIRMSLRKX-AVVLMGKRHNMRRKAIRGLHEENN--PALE 76  
RLAO\_DICDI ----- MSCAS\_SKREKLIFIEKATKLFTTDKMKIVAEADFVGSSULSKIKRSI-----MSCAS\_SKREKNEIEKATKLFTTDKMKIVAEADFVGSSULSKIKRSI----- 76  
Q5ALP0\_DICDI ----- MPREDRATWPSNYFLKELIPLDDPKFCFVYGADNVGSKM0IIRMSLRKX-AVVLMGKRHNMRRKAIRGLHEENN--PALE 76  
RLAO\_PLAF8 ----- MAKLSKQOKKQMYIEKTLSSLIQQSKLIIILVHVNDHSASVKEKSI----- 76  
RLAO\_SULAC ----- MIGLAVTTKKLAKKWWDEVAEETTEKLTAKTITKLIIGRFPAKDLHIEKKKL----- 76  
RLAO\_SULTO ----- MRIMAVITOEQRKLAKKWIEEVKELEKTEKLTAKTITKLIIGRFPAKDLHIEKKKL----- 76  
RLAO\_SULSO ----- MKRLALALQRKVRSNLKEVKELTIELINSNTILLGLEGPAKDLHIEKKKL----- 76  
RLAO\_AERPE ----- MSVVSIVGOMYKQRPIDEPWTLMRLEELSFPSKIRVVFVGRVKKL----- 76  
RLAO\_PYRAK ----- MMLAIGKRRYVYRTPQYARKVKINSEELIICSKVFGWVRLHLSLRIILHEFTRYL----- 76  
RLAO\_METAC ----- MAERHRTHEHIPQWKDEIEENIKELICSIKVFGMWVRIEGILLAKTKIICKRL----- 76  
RLAO\_METMA ----- MAERHRTHEHIPQWKDEIEENIKELICSIKVFGMWVRIEGILLAKTKIICKRL----- 76  
RLAO\_ARCFU ----- MAAVRGS-S-PPEYYKRAVEEIKRMISKDPKMKTRREFRGKX-AEIKVKGKQKLLRFLDRLAEDLGKDALG----- 75  
RLAO\_METKA ----- MAVKRKGQPSGCEPKVAEEWRREVKELKLMDEIEENKLMEVLDLWDRDIPAPOLCEIRAKLRLERDILIRMBRNLMRALEEKDLR-EPE 88  
RLAO\_METTE ----- MAHVAAWKEKEVNLKELIQLKTFVQVWVRLADIPARCLKHMQTLRIDS-ALIRUMKKLISIALEKAGREL-ENV 74  
RLAO\_METTL ----- MITTAESHEHKIAWPWIEEVNLKELIQLKTFVQVWVRLADIPARCLKHMQTLRIDS-ALIRUMKKLISIALEKAGREL-ENV 82  
RLAO\_METVA ----- MIDAKASEHKIAWPWIEEVNLKELIQLSANTVLDIMMVEVPLGEIRDKIR-DOMELKMRNRLIKRRAVEVAEEETGNPEFA 82  
RLAO\_METJA ----- METVKAHAWPVAPPWIEEVKTKLIGSKSPVVAIWMMDVPAULIEIRDKIR-DKVKLMGRNRLIILRALEKAEAAELNPKLIA 81  
RLAO\_PYRAK ----- MAHVAAWKEKEVNLKELIQLKTFVQVWVRLADIPARCLKHMQTLRIDS-ALIRUMKKLISIALEKAGREL-ENV 77  
RLAO\_PYRHO ----- MAHVAAWKEKEVNLKELIQLKTFVQVWVRLADIPARCLKHMQTLRIDS-ALIRUMKKLISIALEKAGREL-ENV 77  
RLAO\_PYRFU ----- MAHVAAWKEKEVNLKELIQLKTFVQVWVRLADIPARCLKHMQTLRIDS-ALIRUMKKLISIALEKAGREL-ENV 77  
RLAO\_PYRK ----- MAHVAAWKEKEVNLKELIQLKTFVQVWVRLADIPARCLKHMQTLRIDS-ALIRUMKKLISIALEKAGREL-ENV 76  
RLAO\_HALMA ----- MSAESERKTETIPWQKQEEDVAVMFIKRSIESEVSGVYVNIAGISLROLSEHRLGGS-AAVRMGRNRLVNRLDGVN----DGF 79  
RLAO\_HALVO ----- MSESERQTETIPWQKQEEDVDFIDESEVSGVYVNIAGISLROLSEHRLGGS-AAVRMGRNRLVNRLDGVN----DGF 79  
RLAO\_HALSA ----- MSAEQRQTETIPWQKQEAEVLDLTSVSVWVNTGIEKQUDLDMRERLIGQ-AALRMRNRLVLALEKAEAEELNPKLIA 79  
RLAO\_THEAC ----- MKEFSQOKKELVNEYITRRIKASRSVALDPIRGEIRAKLRLERDILIRMBRNLMRALEEKDLR-EPE 72  
RLAO\_THEVO ----- MRKRVKKEVNLSELADITSKRAVELDPIRGEIRAKLRLERDILIRMBRNLMRALEEKDLR-EPE 72  
RLAO\_PICTO ----- MTEPROWIDFWKVNENEINSRKVVAALYSKLNEMFKURKNSRDK-AEIRVBRARLRLAIENFK---NNEV 72  
ruler 1.....10.....20.....30.....40.....50.....60.....70.....80.....90.....99

MSA protein sequences

Pfam

# Data

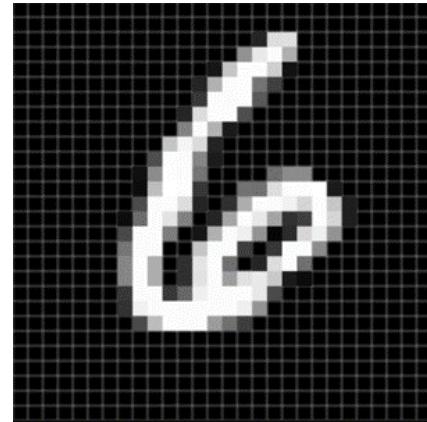
$$\mathcal{D} = \left\{ \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)} \right\}$$

↓

*m*-th entry  $\boldsymbol{x}^{(m)} = \begin{bmatrix} x_1^{(m)} \\ \vdots \\ x_N^{(m)} \end{bmatrix}$

3	8	6	9	6	4	5	3	8	4	5	2	3	8	4	8
1	5	0	5	9	7	4	1	0	3	0	6	2	9	9	4
1	3	6	8	0	7	7	6	8	9	0	3	8	3	7	7
8	4	4	1	2	9	8	1	1	0	6	6	5	0	1	1

M: # of examples in the data set



$N = 28 \times 28$

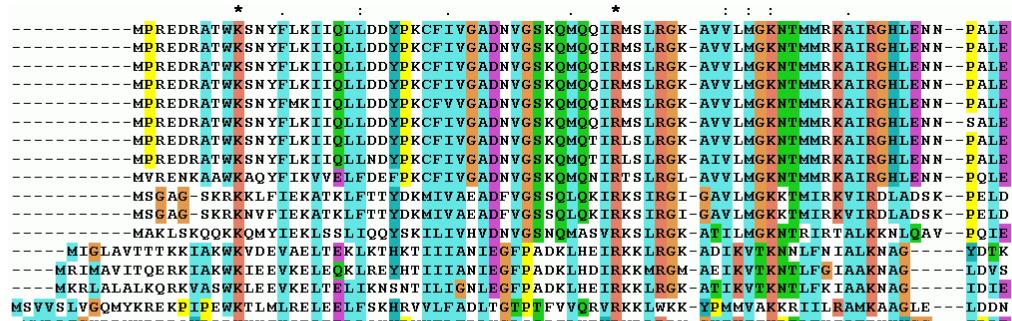
pixels

# Data

$$\mathcal{D} = \left\{ \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)} \right\}$$

↓

*m*-th entry  $\boldsymbol{x}^{(m)} = \begin{bmatrix} x_1^{(m)} \\ \vdots \\ x_N^{(m)} \end{bmatrix}$



The figure shows a sequence logo visualization for a protein family. It consists of multiple protein sequences aligned vertically. Each sequence is composed of amino acids, represented by colored bars. The colors represent the frequency of each amino acid at each position. The sequence logo is used to predict the probability of each amino acid at each position based on the surrounding context.

M: # of sequences in a protein family

GSKOMOQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN-PALE

$$N = L_{MSA} \quad \text{Amino-acids}$$

Goal: Create synthetic sequences

# Data

$$\mathcal{D} = \left\{ \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)} \right\}$$

↓

*m*-th entry     $\boldsymbol{x}^{(m)} = \begin{bmatrix} x_1^{(m)} \\ \vdots \\ x_N^{(m)} \end{bmatrix}$

	$\in \mathbb{R}^N$	continuous
	$\in [0, 1]^N$	binary
	$\in [G, A \dots, N, Q, -]^N$	categorical

# Data distribution

$$\mathcal{D} = \left\{ \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)} \right\}$$

Underlying assumption

i. i. d. realizations of a random variable

$$\boldsymbol{X} \sim P_{\text{data}}$$

(Generally  
unknown)

# Empirical data distribution

$$\mathcal{D} = \left\{ \boldsymbol{x}_d^{(1)}, \dots, \boldsymbol{x}_d^{(M)} \right\}$$

Underlying assumption

i. i. d. realizations of a random variable

$$\boldsymbol{X}_d \sim P_{\text{data}}$$

(Generally  
unknown)

Empirical distribution

$$p_{\mathcal{D}}(\boldsymbol{x}) = \frac{1}{M} \sum_{m=1}^M \delta \left( \boldsymbol{x} - \boldsymbol{x}_d^{(m)} \right) \xrightarrow{\text{Large } M} p_{\text{data}}(\boldsymbol{x})$$

# Empirical data distribution

$$\mathcal{D} = \left\{ \boldsymbol{x}_d^{(1)}, \dots, \boldsymbol{x}_d^{(M)} \right\}$$

Underlying assumption

i. i. d. realizations of a random variable

$$\boldsymbol{X}_d \sim P_{\text{data}}$$

(Generally  
unknown)

Empirical distribution

$$p_{\mathcal{D}}(\boldsymbol{x}) = \frac{1}{M} \sum_{m=1}^M \delta \left( \boldsymbol{x} - \boldsymbol{x}_d^{(m)} \right) \xrightarrow{\text{Large } M} p_{\text{data}}(\boldsymbol{x})$$

Generative models :  $p_{\boldsymbol{\theta}}(\boldsymbol{x})$        $\boldsymbol{\theta}$



# Energy-based models

# Energy based models (EBMs)

Hinton, Hopfield, LeCun, Bengio

<i>Empirical</i>	<i>Model</i>
$p_{\mathcal{D}}(\mathbf{x}) \sim p_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{e^{-E_{\boldsymbol{\theta}}(\mathbf{x})}}{Z_{\boldsymbol{\theta}}}$	

Gibbs-Boltzmann distribution

$E_{\boldsymbol{\theta}}(\mathbf{x})$  energy function

$$Z_{\boldsymbol{\theta}} = \int d\mathbf{x} e^{-E_{\boldsymbol{\theta}}(\mathbf{x})} \quad \text{Partition function}$$

**Learning** : adjust the parameters  $\boldsymbol{\theta}$  so that the dataset configurations are **typical** configurations of the model.

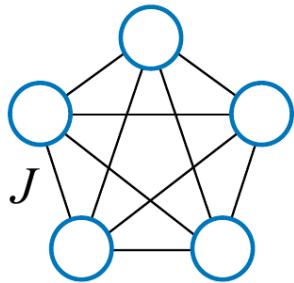
# Energy based models (EBMs)

Hinton, Hopfield, LeCun, Bengio

Boltzmann Machines (Ising/Hopfield/Potts models)

- Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). *A learning algorithm for Boltzmann machines*. Cognitive science, 9(1), 147-169.

$$E_{\theta}(\mathbf{x})$$



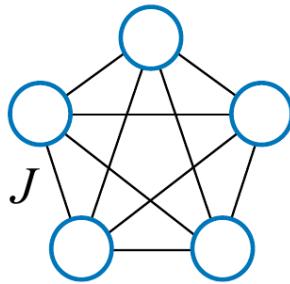
$$E_{J,h}(\mathbf{x}) = -\mathbf{x}^\top \mathbf{J} \mathbf{x} - \mathbf{h}^\top \mathbf{x}$$

Pairwise interactions

# Energy based models (EBMs)

Hinton, Hopfield, LeCun, Bengio

- Ising/Hopfield/Potts models

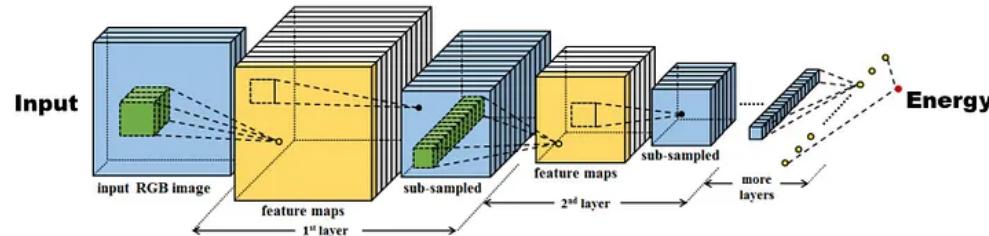


$$E_{J,h}(x) = -x^\top J x - h^\top x$$

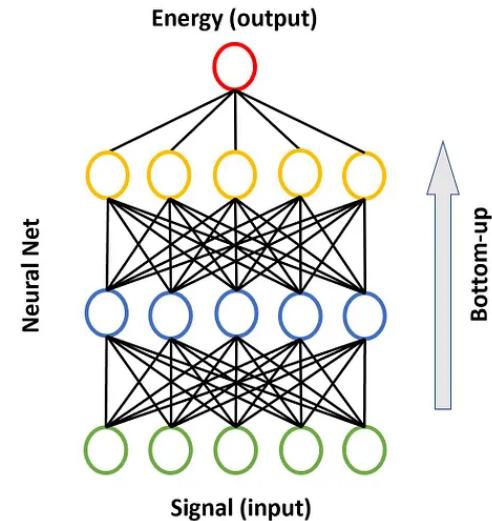
- Generative ConvNets

- LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., & Huang, F. (2006). *A tutorial on energy-based learning*.  
- Xie, J., Lu, Y., Zhu, S. C., & Wu, Y. (2016). *A theory of generative convnet*.

A feedforward ConvNet that maps the input signal to an energy or a score



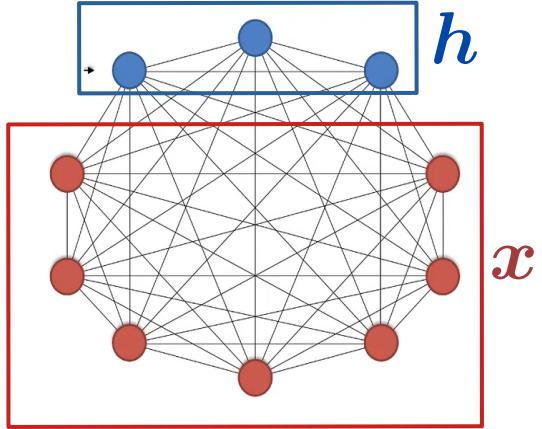
$$E_{\theta}(x)$$



# Models with hidden variables

- Boltzmann Machines

- Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). *A learning algorithm for Boltzmann machines*. Cognitive science, 9(1), 147-169.

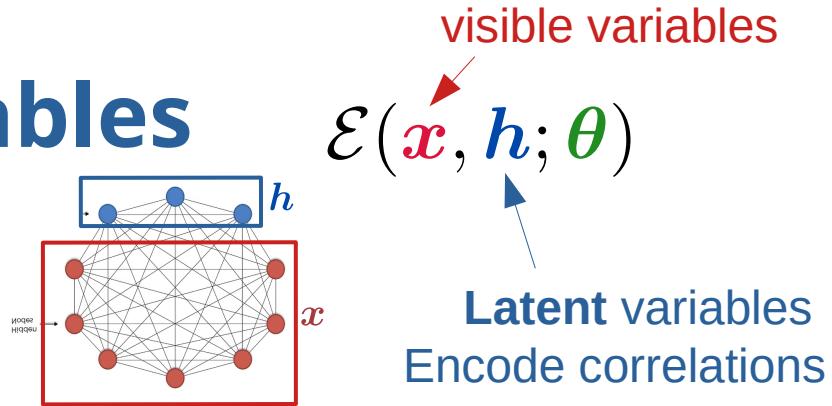


visible variables  
 $\mathcal{E}(\textcolor{red}{x}, \textcolor{blue}{h}; \theta)$   
Latent variables  
Encode correlations

# Models with hidden variables

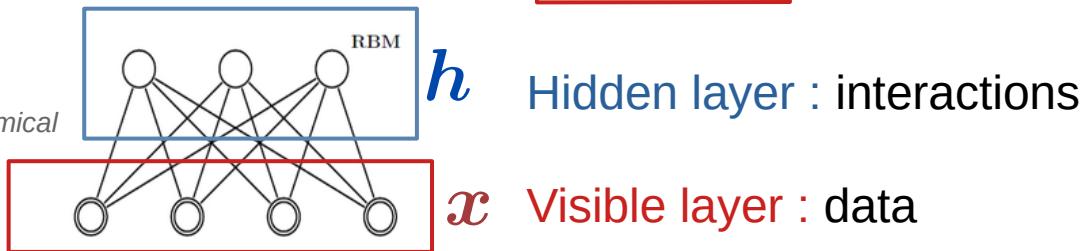
- Boltzmann Machines

- Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). *A learning algorithm for Boltzmann machines*. Cognitive science, 9(1), 147-169.



- Restricted Boltzmann Machine

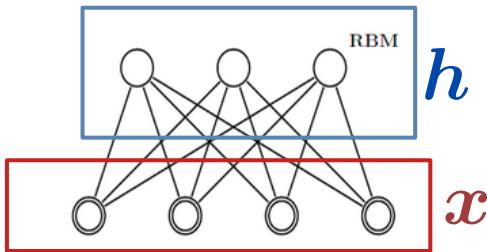
- Smolensky, P. (1986). *Information processing in dynamical systems: Foundations of harmony theory*.



# Models with hidden variables

- Restricted Boltzmann Machine

- Smolensky, P. (1986). *Information processing in dynamical systems: Foundations of harmony theory.*



visible variables  
Latent variables  
Encode correlations

$$\mathcal{E}(\mathbf{x}, \mathbf{h}; \theta)$$

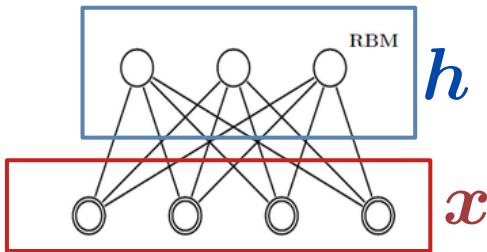
$$\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h}) = -\mathbf{x}^\top W \mathbf{h} - \boldsymbol{\zeta}^\top \mathbf{x} - \boldsymbol{\eta}^\top \mathbf{h} \quad \theta = \{W, \boldsymbol{\zeta}, \boldsymbol{\eta}\}$$

$$p_{\theta}(\mathbf{x}) = \frac{e^{-E_{\theta}(\mathbf{x})}}{Z_{\theta}} = \frac{\sum_{\mathbf{h}} e^{-\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h})}}{Z_{\theta}} = \frac{e^{\sum_i x_i \zeta_i}}{Z_{\theta}} \prod_{a=1}^{N_h} \sum_{\mathbf{h}_a=0}^1 e^{\sum_i x_i W_{ia} h_a + \eta_a h_a}$$

# Models with hidden variables

- Restricted Boltzmann Machine

- Smolensky, P. (1986). *Information processing in dynamical systems: Foundations of harmony theory.*



visible variables  
 $\mathcal{E}(\mathbf{x}, \mathbf{h}; \theta)$   
Latent variables  
Encode correlations

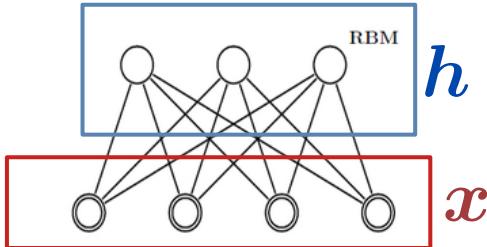
$$\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h}) = -\mathbf{x}^\top W \mathbf{h} - \boldsymbol{\zeta}^\top \mathbf{x} - \boldsymbol{\eta}^\top \mathbf{h} \quad \theta = \{W, \boldsymbol{\zeta}, \boldsymbol{\eta}\}$$

$$\begin{aligned} p_{\theta}(\mathbf{x}) &= \frac{e^{-E_{\theta}(\mathbf{x})}}{Z_{\theta}} = \frac{\sum_{\mathbf{h}} e^{-\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h})}}{Z_{\theta}} = \frac{e^{\sum_i x_i \zeta_i}}{Z_{\theta}} \prod_{a=1}^{N_h} \sum_{\mathbf{h}_a=0}^1 e^{\sum_i x_i W_{ia} \mathbf{h}_a + \eta_a \mathbf{h}_a} \\ &= \frac{e^{\sum_i x_i \zeta_i}}{Z_{\theta}} \prod_{a=1}^{N_h} \left( 1 + e^{\sum_i x_i W_{ia} + \eta_a} \right) \end{aligned}$$

# Models with hidden variables

- Restricted Boltzmann Machine

- Smolensky, P. (1986). *Information processing in dynamical systems: Foundations of harmony theory.*



visible variables  
 $\mathcal{E}(\textcolor{red}{x}, \textcolor{blue}{h}; \theta)$   
 Latent variables  
 Encode correlations

$$\mathcal{E}_\theta(\textcolor{red}{x}, \textcolor{blue}{h}) = -\textcolor{red}{x}^\top W \textcolor{blue}{h} - \zeta^\top \textcolor{red}{x} - \eta^\top \textcolor{blue}{h}$$

$$\theta = \{W, \zeta, \eta\}$$

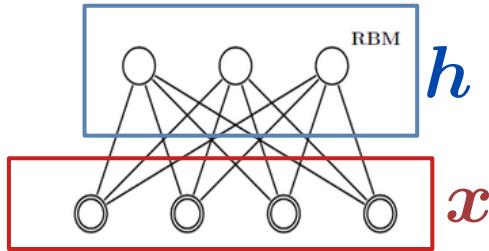
$$p_\theta(\textcolor{red}{x}) = \frac{e^{-E_\theta(\textcolor{red}{x})}}{Z_\theta} = \frac{\sum_{\textcolor{blue}{h}} e^{-\mathcal{E}_\theta(\textcolor{red}{x}, \textcolor{blue}{h})}}{Z_\theta} = \frac{e^{\sum_i x_i \zeta_i}}{Z_\theta} \prod_{a=1}^{N_h} \sum_{\textcolor{blue}{h}_a=0}^1 e^{\sum_i x_i W_{ia} \textcolor{blue}{h}_a + \eta_a \textcolor{blue}{h}_a}$$

$$= \frac{e^{\sum_i x_i \zeta_i}}{Z_\theta} \prod_{a=1}^{N_h} \left( 1 + e^{\sum_i x_i W_{ia} + \eta_a} \right) \Rightarrow E_\theta(\textcolor{red}{x}) = - \sum_i x_i \zeta_i - \sum_{a=1}^{N_h} \log \left( 1 + e^{\sum_i \textcolor{red}{x}_i W_{ia} + \eta_a} \right)$$

# Models with hidden variables

- Restricted Boltzmann Machine

- Smolensky, P. (1986). *Information processing in dynamical systems: Foundations of harmony theory.*



visible variables  
 $\mathcal{E}(\textcolor{red}{x}, \textcolor{blue}{h}; \theta)$   
 Latent variables  
 Encode correlations

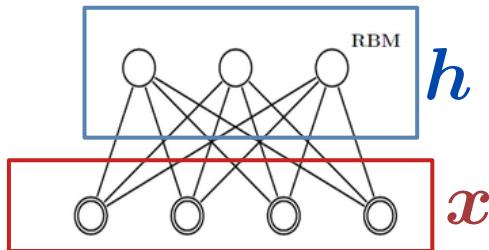
$$\mathcal{E}_\theta(\textcolor{red}{x}, \textcolor{blue}{h}) = -\textcolor{red}{x}^\top W \textcolor{blue}{h} - \zeta^\top \textcolor{red}{x} - \eta^\top \textcolor{blue}{h} \quad \theta = \{W, \zeta, \eta\}$$

$$\begin{aligned}
 p_\theta(\textcolor{red}{x}) &= \frac{e^{-E_\theta(\textcolor{red}{x})}}{Z_\theta} = \frac{\sum_{\textcolor{blue}{h}} e^{-\mathcal{E}_\theta(\textcolor{red}{x}, \textcolor{blue}{h})}}{Z_\theta} = \frac{e^{\sum_i x_i \zeta_i}}{Z_\theta} \prod_{a=1}^{N_h} \sum_{\textcolor{blue}{h}_a=0}^1 e^{\sum_i x_i W_{ia} \textcolor{blue}{h}_a + \eta_a \textcolor{blue}{h}_a} \\
 &= \frac{e^{\sum_i x_i \zeta_i}}{Z_\theta} \prod_{a=1}^{N_h} \left( 1 + e^{\sum_i x_i W_{ia} + \eta_a} \right) \Rightarrow E_\theta(\textcolor{red}{x}) = -\sum_i x_i \zeta_i - \sum_{a=1}^{N_h} \log \left( 1 + e^{\sum_i \textcolor{red}{x}_i W_{ia} + \eta_a} \right) \\
 &\Rightarrow E_\theta(\textcolor{red}{x}) = -\sum_i \textcolor{cyan}{h}_i x_i - \sum_{ij} \textcolor{cyan}{J}_{ij}^{(2)} x_i x_j - \sum_{ijk} \textcolor{cyan}{J}_{ijk}^{(3)} x_i x_j x_k - \sum_{ijkl} \textcolor{cyan}{J}_{ijkl}^{(4)} x_i x_j x_k x_l + \dots
 \end{aligned}$$

# Models with hidden variables

- Restricted Boltzmann Machine

- Smolensky, P. (1986). *Information processing in dynamical systems: Foundations of harmony theory.*



visible variables  
 $\mathcal{E}(\textcolor{red}{x}, \textcolor{blue}{h}; \theta)$   
 Latent variables  
 Encode correlations

$$\mathcal{E}_\theta(\textcolor{red}{x}, \textcolor{blue}{h}) = -\textcolor{red}{x}^\top W \textcolor{blue}{h} - \zeta^\top \textcolor{red}{x} - n^\top \textcolor{blue}{h} \quad \theta = \{W, \zeta, n\}$$

The marginal energy for the RBM encode high order interactions! → **Universal approximator**

$$p_\theta(\textcolor{red}{x}) = \frac{e^{-E_\theta(\textcolor{red}{x})}}{Z_\theta} = \frac{\sum_{\textcolor{blue}{h}}}{Z_\theta}$$

$$= \frac{e^{\sum_i x_i \zeta_i}}{Z_\theta} \prod_{a=1}^{N_h} \left( \frac{1}{1 + e^{\sum_i x_i \zeta_i + \eta_a}} \right) \Rightarrow E_\theta(\textcolor{red}{x}) = -\sum_i x_i \zeta_i - \sum_{a=1} \log \left( \frac{1}{1 + e^{\sum_i x_i \zeta_i + \eta_a}} \right)$$

Le Roux and Bengio. Neural computation (2008)

$$\Rightarrow E_\theta(\textcolor{red}{x}) = -\sum_i h_i x_i - \sum_{ij} J_{ij}^{(2)} x_i x_j - \sum_{ijk} J_{ijk}^{(3)} x_i x_j x_k - \sum_{ijkl} J_{ijkl}^{(4)} x_i x_j x_k x_l + \dots$$

# Models with hidden variables

- Boltzmann Machines (Ising/Hopfield/Potts models)

- Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). *A learning algorithm for Boltzmann machines*. Cognitive science, 9(1), 147-169.

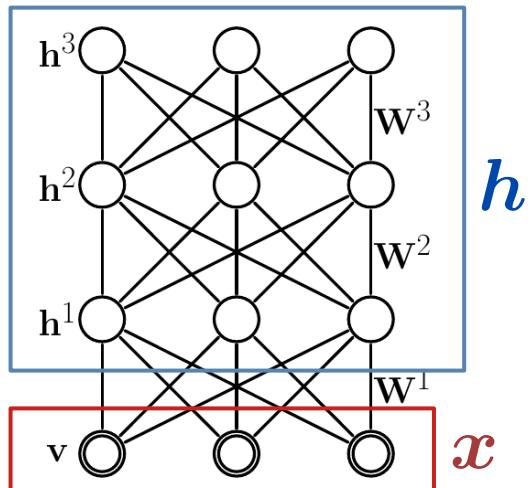
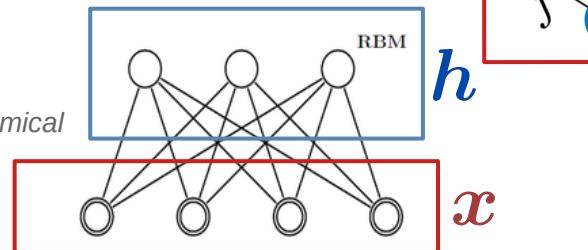
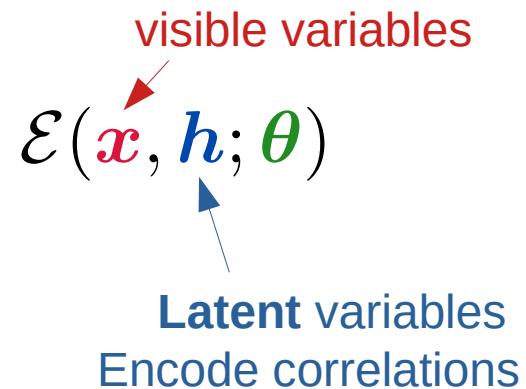
- Restricted Boltzmann Machine

- Smolensky, P. (1986). *Information processing in dynamical systems: Foundations of harmony theory*.

- Deep Boltzmann Machines

- Ruslan Salakhutdinov, Geoffrey Hinton (2009) Deep Boltzmann Machines.

- Bengio, Y. (2009). *Learning deep architectures for AI*.



$$p_{\theta}(x) = \frac{\sum_h e^{-\mathcal{E}_{\theta}(x, h)}}{Z_{\theta}}$$

$$= \frac{e^{-E_{\theta}(x)}}{Z_{\theta}} \sim p_{\mathcal{D}}(x)$$



# Training procedure

3	8	4	5	2	3	8	4	8
1	0	3	0	6	2	9	9	4
6	8	9	0	3	8	3	7	1
1	1	0	6	4	5	0	1	1

# Training procedure

Goal of the training:

$$p_{\mathcal{D}}(\mathbf{x}) \sim p_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{e^{-E_{\boldsymbol{\theta}}(\mathbf{x})}}{Z_{\boldsymbol{\theta}}}$$

$$p_{\mathcal{D}}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^M \delta \left( \mathbf{x} - \mathbf{x}^{(m)} \right)$$

# Training procedure

Goal of the training:

$$p_{\mathcal{D}}(\mathbf{x}) \sim p_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{e^{-E_{\boldsymbol{\theta}}(\mathbf{x})}}{Z_{\boldsymbol{\theta}}}$$

Minimize  
Kullback-Leibler (KL)  
divergence

$$\begin{aligned} D_{\text{KL}}(p_{\mathcal{D}} || p_{\boldsymbol{\theta}}) &= \int d\mathbf{x} p_{\mathcal{D}}(\mathbf{x}) \log \frac{p_{\mathcal{D}}(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})} \\ &= \underbrace{\int d\mathbf{x} p_{\mathcal{D}}(\mathbf{x}) \log p_{\mathcal{D}}(\mathbf{x})}_{\text{Constant}} - \underbrace{\int d\mathbf{x} p_{\mathcal{D}}(\mathbf{x}) \log p_{\boldsymbol{\theta}}(\mathbf{x})}_{\text{Variable}} \end{aligned}$$

# Training procedure

$$p_{\mathcal{D}}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^M \delta(\mathbf{x} - \mathbf{x}^{(m)})$$

Goal of the training:

$$p_{\mathcal{D}}(\mathbf{x}) \sim p_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{e^{-E_{\boldsymbol{\theta}}(\mathbf{x})}}{Z_{\boldsymbol{\theta}}}$$

Minimize  
Kullback-Leibler (KL)  
divergence

$$\begin{aligned} D_{\text{KL}}(p_{\mathcal{D}} || p_{\boldsymbol{\theta}}) &= \int d\mathbf{x} p_{\mathcal{D}}(\mathbf{x}) \log \frac{p_{\mathcal{D}}(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})} \\ &= \int d\mathbf{x} p_{\mathcal{D}}(\mathbf{x}) \log p_{\mathcal{D}}(\mathbf{x}) - \underbrace{\int d\mathbf{x} p_{\mathcal{D}}(\mathbf{x}) \log p_{\boldsymbol{\theta}}(\mathbf{x})}_{\text{Constant}} \end{aligned}$$

log-likelihood

$$-\frac{1}{M} \sum_{m=1}^M \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(m)}) = -\frac{1}{M} \log \prod_{m=1}^M p_{\boldsymbol{\theta}}(\mathbf{x}^{(m)}) = -\frac{1}{M} \log L(\mathcal{D} | \boldsymbol{\theta})$$

# Training procedure

$$p_{\mathcal{D}}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^M \delta(\mathbf{x} - \mathbf{x}^{(m)})$$

Goal of the training:

$$p_{\mathcal{D}}(\mathbf{x}) \sim p_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{e^{-E_{\boldsymbol{\theta}}(\mathbf{x})}}{Z_{\boldsymbol{\theta}}}$$

Minimize

Kullback-Leibler divergence

divergence

Maximize

The log-likelihood

$$\log L(\mathcal{D}|\boldsymbol{\theta}) \equiv \mathcal{L}(\mathcal{D}|\boldsymbol{\theta})$$

$$= \int d\mathbf{x} p_{\mathcal{D}}(\mathbf{x}) \log p_{\mathcal{D}}(\mathbf{x}) - \left( \int d\mathbf{x} p_{\mathcal{D}}(\mathbf{x}) \log p_{\boldsymbol{\theta}}(\mathbf{x}) \right)$$

Constant

$$-\frac{1}{M} \sum_{m=1}^M \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(m)}) = -\frac{1}{M} \log \prod_{m=1}^M p_{\boldsymbol{\theta}}(\mathbf{x}^{(m)}) = -\frac{1}{M} \log L(\mathcal{D}|\boldsymbol{\theta})$$

# Training procedure

$$p_{\mathcal{D}}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^M \delta \left( \mathbf{x} - \mathbf{x}_d^{(m)} \right)$$

Goal of the training:

$$p_{\mathcal{D}}(\mathbf{x}) \sim p_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{e^{-E_{\boldsymbol{\theta}}(\mathbf{x})}}{Z_{\boldsymbol{\theta}}}$$

Minimize

Kullback-Leibler divergence

divergence

Maximize

The log-likelihood

$$\log L(\mathcal{D}|\boldsymbol{\theta}) \equiv \mathcal{L}(\mathcal{D}|\boldsymbol{\theta})$$

Recall Bayes-Theorem

$$p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = \underbrace{p(\boldsymbol{\theta}|\mathcal{D})}_{\text{likelihood}} \underbrace{p(\mathcal{D})}_{\text{posterior}}$$

$$= \int d\mathbf{x} p_{\mathcal{D}}(\mathbf{x}) \log p_{\mathcal{D}}(\mathbf{x}) - \underbrace{\left( \int d\mathbf{x} p_{\mathcal{D}}(\mathbf{x}) \log p_{\boldsymbol{\theta}}(\mathbf{x}) \right)}_{\text{constant}}$$

constant

$$-\frac{1}{M} \log \prod_{m=1}^M p_{\boldsymbol{\theta}}(\mathbf{x}^{(m)}) = -\frac{1}{M} \log L(\mathcal{D}|\boldsymbol{\theta})$$

# Log-likelihood maximization

$$\mathcal{L}(\mathcal{D}|\boldsymbol{\theta}) = \sum_{m=1}^M \log p_{\boldsymbol{\theta}}(x = x^{(m)})$$

$$p_{\boldsymbol{\theta}}(x) = \frac{e^{-E_{\boldsymbol{\theta}}(x)}}{Z_{\boldsymbol{\theta}}}$$

# Log-likelihood maximization

$$\mathcal{L}(\mathcal{D}|\theta) = \sum_{m=1}^M \log p_\theta(x = x^{(m)})$$

$$p_\theta(x) = \frac{e^{-E_\theta(x)}}{Z_\theta}$$

$$\mathcal{L}(\mathcal{D}|\theta) = \langle \log p_\theta(x) \rangle_{p_{\mathcal{D}}} = \langle -E_\theta(x) \rangle_{p_{\mathcal{D}}}$$

Partition function

$$\log Z_\theta$$

$$Z_\theta = \sum_{\{x\}} e^{-E_\theta(x)}$$

If  $x_i$  binary  $\rightarrow 2^N$

Intractable

# Log-likelihood maximization

$$\mathcal{L}(\mathcal{D}|\boldsymbol{\theta}) = \sum_{m=1}^M \log p_{\boldsymbol{\theta}}(x = x^{(m)})$$

$$p_{\boldsymbol{\theta}}(x) = \frac{e^{-E_{\boldsymbol{\theta}}(x)}}{Z_{\boldsymbol{\theta}}}$$

$$\mathcal{L}(\mathcal{D}|\boldsymbol{\theta}) = \langle \log p_{\boldsymbol{\theta}}(x) \rangle_{p_{\mathcal{D}}} = \langle -E_{\boldsymbol{\theta}}(x) \rangle_{p_{\mathcal{D}}}$$

Partition function

$$\log Z_{\boldsymbol{\theta}}$$

$$Z_{\boldsymbol{\theta}} = \sum_{\{x\}} e^{-E_{\boldsymbol{\theta}}(x)}$$

(Stochastic) gradient **ascent**

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}$$

$$\theta_i^{(t+1)} \leftarrow \theta_i^t + \gamma \left. \frac{\partial \mathcal{L}}{\partial \theta_i} \right|_{\theta=\theta_i^{(t)}}$$

If  $x_i$  binary  $\rightarrow 2^N$

Intractable

# Log-likelihood maximization

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = \left\langle -\frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\mathcal{D}}} - \frac{\partial \log Z}{\partial \theta_i}$$

$$\mathcal{L}(\mathcal{D}|\boldsymbol{\theta}) = \langle \log p_{\boldsymbol{\theta}}(\mathbf{x}) \rangle_{p_{\mathcal{D}}} = \langle -E_{\boldsymbol{\theta}}(\mathbf{x}) \rangle_{p_{\mathcal{D}}}$$

Partition function

$$-\log Z_{\boldsymbol{\theta}}$$

(Stochastic) gradient **ascent**

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}$$

$$\theta_i^{(t+1)} \leftarrow \theta_i^t + \gamma \left. \frac{\partial \mathcal{L}}{\partial \theta_i} \right|_{\theta=\theta_i^{(t)}}$$

# Log-likelihood maximization

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = \left\langle -\frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\mathcal{D}}} - \frac{\partial \log Z}{\partial \theta_i}$$

$$\begin{aligned}\frac{\partial \log Z}{\partial \theta_i} &= \sum_{\{\mathbf{x}\}} \frac{e^{-E(\mathbf{x})}}{Z} \frac{\partial E(\mathbf{x})}{\partial \theta_i} \\ &= \left\langle \frac{\partial E(\mathbf{x})}{\partial \theta_i} \right\rangle_{p_{\boldsymbol{\theta}}(\mathbf{x})}\end{aligned}$$

$$\mathcal{L}(\mathcal{D}|\boldsymbol{\theta}) = \langle \log p_{\boldsymbol{\theta}}(\mathbf{x}) \rangle_{p_{\mathcal{D}}} = \langle -E_{\boldsymbol{\theta}}(\mathbf{x}) \rangle_{p_{\mathcal{D}}} + \log Z_{\boldsymbol{\theta}}$$

(Stochastic) gradient **ascent**

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}$$

$$\theta_i^{(t+1)} \leftarrow \theta_i^t + \gamma \frac{\partial \mathcal{L}}{\partial \theta_i} \Big|_{\theta=\theta_i^{(t)}}$$

$$\nabla \mathcal{L}_{\boldsymbol{\theta}} = \underbrace{\langle -\nabla E_{\boldsymbol{\theta}} \rangle_{p_{\mathcal{D}}}}_{\text{data}} - \underbrace{\langle -\nabla E_{\boldsymbol{\theta}} \rangle_{p_{\boldsymbol{\theta}}}}_{\text{model}}$$

# Log-likelihood maximization

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = \left\langle -\frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\mathcal{D}}}$$

$$-\frac{\frac{\partial \log Z}{\partial \theta_i}}{\partial \theta_i}$$

$$\overline{\nabla E}_{\theta}^{\mathcal{D}}$$

$$\begin{aligned}\frac{\partial \log Z}{\partial \theta_i} &= \sum_{\{x\}} \frac{e^{-E(x)}}{Z} \frac{\partial E(x)}{\partial \theta_i} \\ &= \left\langle \frac{\partial E(x)}{\partial \theta_i} \right\rangle_{p_{\theta}(x)}\end{aligned}$$

$$p_{\mathcal{D}}(x) = \frac{1}{M} \sum_{m=1}^M \delta(x - x_d^{(m)})$$

(Stochastic) gradient **ascent**

$$\nabla_{\theta} \mathcal{L}$$

$$\theta_i^{(t+1)} \leftarrow \theta_i^t + \gamma \frac{\partial \mathcal{L}}{\partial \theta_i} \Big|_{\theta=\theta_i^{(t)}}$$

$$\nabla \mathcal{L}_{\theta} = \underbrace{\left\langle -\nabla E_{\theta} \right\rangle_{p_{\mathcal{D}}}}_{\text{data}} - \underbrace{\left\langle -\nabla E_{\theta} \right\rangle_{p_{\theta}}}_{\text{model}}$$

Easy !

# Log-likelihood maximization

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = \left\langle -\frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\mathcal{D}}}$$

$$\frac{\partial \log Z}{\partial \theta_i}$$

$$\begin{aligned}\frac{\partial \log Z}{\partial \theta_i} &= \sum_{\{x\}} \frac{e^{-E(x)}}{Z} \frac{\partial E(x)}{\partial \theta_i} \\ &= \left\langle \frac{\partial E(x)}{\partial \theta_i} \right\rangle \\ p_{\theta}(x) &= \frac{e^{-E_{\theta}(x)}}{Z_{\theta}}\end{aligned}$$

$$p_{\mathcal{D}}(x) = \frac{1}{M} \sum_{m=1}^M \delta(x - x_d^{(m)})$$

MCMC sampling

(Stochastic) gradient **ascent**

$$\nabla_{\theta} \mathcal{L}$$

$$\theta_i^{(t+1)} \leftarrow \theta_i^t + \gamma \frac{\partial \mathcal{L}}{\partial \theta_i} \Big|_{\theta=\theta_i^{(t)}}$$

$$\nabla \mathcal{L}_{\theta} = \left\langle -\nabla E_{\theta} \right\rangle_{p_{\mathcal{D}}}$$

**data**

$$\left\langle -\nabla E_{\theta} \right\rangle_{p_{\theta}}$$

**model**

**Easy !**

**Hard !**

# Log-likelihood

Every time  
we want to  
update the  
parameters

$$\boldsymbol{x}_{\text{gen}}^{(m)} \quad m = 1, \dots, n_{\text{chains}}$$

$\boldsymbol{X}_{\text{gen}} \sim P_{\boldsymbol{\theta}}$  Via a Markov Chain Monte Carlo process

$$\langle -\nabla E_{\boldsymbol{\theta}} \rangle_{p_{\boldsymbol{\theta}}} \approx \frac{1}{n_{\text{chains}}} \sum_{m=1}^{n_{\text{chains}}} \nabla E(\boldsymbol{x}_{\text{gen}}^{(m)})$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

$$p_{\mathcal{D}}(\boldsymbol{x}) = \frac{1}{M} \sum_{m=1}^M \delta(\boldsymbol{x} - \boldsymbol{x}_d^{(m)})$$

MCMC sampling

(Stochastic) gradient ascent

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}$$

$$\theta_i^{(t+1)} \leftarrow \theta_i^t + \gamma \frac{\partial \mathcal{L}}{\partial \theta_i} \Big|_{\theta=\theta_i^{(t)}}$$

$$\nabla \mathcal{L}_{\boldsymbol{\theta}} = \underbrace{\langle -\nabla E_{\boldsymbol{\theta}} \rangle_{p_{\mathcal{D}}}}_{\text{data}} - \underbrace{\langle -\nabla E_{\boldsymbol{\theta}} \rangle_{p_{\boldsymbol{\theta}}}}_{\text{model}}$$

Easy !

39 / 69  
Hard !

# Log-likelihood

Every time  
we want to  
update the  
parameters

$$\boldsymbol{x}_{\text{gen}}^{(m)} \quad m = 1, \dots, n_{\text{chains}}$$

$\boldsymbol{X}_{\text{gen}} \sim P_{\boldsymbol{\theta}}$  Via a Markov Chain Monte Carlo process

$$\langle -\nabla E_{\boldsymbol{\theta}} \rangle_{p_{\boldsymbol{\theta}}} \approx \frac{1}{n_{\text{chains}}} \sum_{m=1}^{n_{\text{chains}}} \nabla E(\boldsymbol{x}_{\text{gen}}^{(m)})$$

Origin of all the difficulties ! → 3<sup>rd</sup> lecture

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

$$p_{\mathcal{D}}(\boldsymbol{x}) = \frac{1}{M} \sum_{m=1}^M \delta\left(\boldsymbol{x} - \boldsymbol{x}_d^{(m)}\right)$$

MCMC sampling

(Stochastic) gradient ascent

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}$$

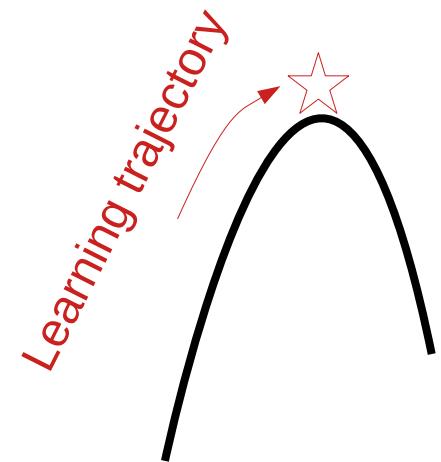
$$\theta_i^{(t+1)} \leftarrow \theta_i^t + \gamma \frac{\partial \mathcal{L}}{\partial \theta_i} \Big|_{\theta=\theta_i^{(t)}}$$

$$\nabla \mathcal{L}_{\boldsymbol{\theta}} = \underbrace{\langle -\nabla E_{\boldsymbol{\theta}} \rangle_{p_{\mathcal{D}}}}_{\text{data}} - \underbrace{\langle -\nabla E_{\boldsymbol{\theta}} \rangle_{p_{\boldsymbol{\theta}}}}_{\text{model}}$$

Easy !

40 / 69  
Hard !

# On the gradient ascent



$$\theta(t + \tau) \leftarrow \theta(t) + \gamma \nabla \mathcal{L}(t)$$

Update rule:

$$\nabla \mathcal{L}_{\theta} = \langle -\nabla E_{\theta} \rangle_{p_{\mathcal{D}}} - \langle -\nabla E_{\theta} \rangle_{p_{\theta}}$$

# On the gradient ascent

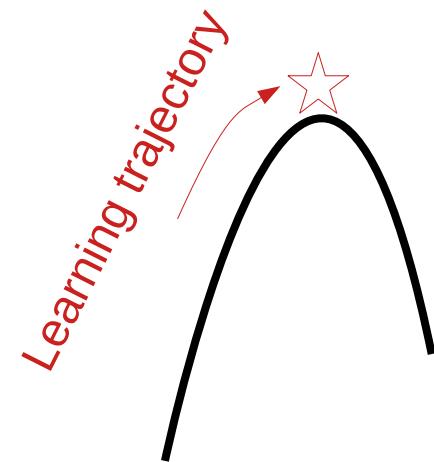
★ Fixed point :  $\nabla \mathcal{L}_{\theta} = 0$

$$\left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\mathcal{D}}} = \left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\theta}} \quad \forall \theta_i$$

Moment matching statistics

Update rule:

$$\theta(t + t) \leftarrow \theta(t) + \gamma \nabla \mathcal{L}(t)$$



$$\nabla \mathcal{L}_{\theta} = \left\langle -\nabla E_{\theta} \right\rangle_{p_{\mathcal{D}}} - \left\langle -\nabla E_{\theta} \right\rangle_{p_{\theta}}$$

# On the gradient ascent

$$f_{\theta_i}(x, \boldsymbol{\theta}) \equiv \frac{\partial E_{\boldsymbol{\theta}}(x)}{\partial \theta_i}$$

★ Fixed point :  $\nabla \mathcal{L}_{\boldsymbol{\theta}} = 0$

$$\left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\mathcal{D}}} = \left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\boldsymbol{\theta}}} \quad \forall \theta_i$$

Moment matching statistics

Hessian matrix

$$H_{ij}(\boldsymbol{\theta}) \equiv \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} = \left\langle \frac{\partial f_{\theta_j}(x)}{\partial \theta_i} \right\rangle_{p_{\boldsymbol{\theta}}} - \left\langle \frac{\partial f_{\theta_j}(x)}{\partial \theta_i} \right\rangle_{p_{\mathcal{D}}} - \left\langle f_{\theta_i}(x, \boldsymbol{\theta}) f_{\theta_j}(x, \boldsymbol{\theta}) \right\rangle_{p_{\boldsymbol{\theta}}} + \left\langle f_{\theta_i}(x, \boldsymbol{\theta}) \right\rangle_{p_{\boldsymbol{\theta}}} \left\langle f_{\theta_j}(x, \boldsymbol{\theta}) \right\rangle_{p_{\boldsymbol{\theta}}}$$

Update rule:

$$\nabla \mathcal{L}_{\boldsymbol{\theta}} = \left\langle -\nabla E_{\boldsymbol{\theta}} \right\rangle_{p_{\mathcal{D}}} - \left\langle -\nabla E_{\boldsymbol{\theta}} \right\rangle_{p_{\boldsymbol{\theta}}}$$

# On the gradient ascent

$$f_{\theta_i}(x, \theta) \equiv \frac{\partial E_{\theta}(x)}{\partial \theta_i}$$

★ Fixed point :  $\nabla \mathcal{L}_{\theta} = 0$

$$\left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\mathcal{D}}} = \left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\theta}} \quad \forall \theta_i$$

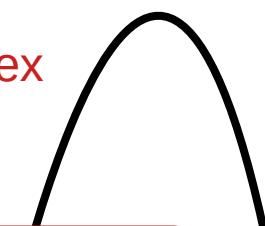
Moment matching statistics

Update rule:

Hessian matrix

$$H_{ij}(\theta) \equiv \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta_i \partial \theta_j} = \left\langle \frac{\partial f_{\theta_j}(x)}{\partial \theta_i} \right\rangle_{p_{\theta}} - \left\langle \frac{\partial f_{\theta_j}(x)}{\partial \theta_i} \right\rangle_{p_{\mathcal{D}}} \\ - \left\langle f_{\theta_i}(x, \theta) f_{\theta_j}(x, \theta) \right\rangle_{p_{\theta}} + \left\langle f_{\theta_i}(x, \theta) \right\rangle_{p_{\theta}} \left\langle f_{\theta_j}(x, \theta) \right\rangle_{p_{\theta}}$$

Semi negative definite  $\Rightarrow$  convex



$$\nabla \mathcal{L}_{\theta} = \left\langle -\nabla E_{\theta} \right\rangle_{p_{\mathcal{D}}} - \left\langle -\nabla E_{\theta} \right\rangle_{p_{\theta}}$$

# Example 1: Boltzmann Machine

★ Fixed point :  $\nabla \mathcal{L}_{\theta} = 0$

$$\left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\mathcal{D}}} = \left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\theta}} \quad \forall \theta_i$$

Moment matching statistics

$$\frac{\partial \mathcal{L}}{\partial J_{ij}} = \langle S_i S_j \rangle_{p_{\mathcal{D}}} - \langle S_i S_j \rangle_{p_{\theta}}$$

$$\frac{\partial \mathcal{L}}{\partial h_i} = \langle S_i \rangle_{p_{\mathcal{D}}} - \langle S_i \rangle_{p_{\theta}}$$

Ising-like model

$$E_{J, h}(S) = - \sum_{ij} J_{ij} S_i S_j - \sum_i h_i S_i$$

$$\frac{\partial E}{\partial J_{ij}} = -S_i S_j \quad \frac{\partial E}{\partial h_i} = -S_i$$

# Example 1: Boltzmann Machine

★ Fixed point :

$$\nabla \mathcal{L}_{\theta} = 0$$

$$\left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\mathcal{D}}} = \left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\theta}} \quad \forall \theta_i$$

Moment matching statistics

Ising-like model

$$E_{J,h}(S) = - \sum_{ij} J_{ij} S_i S_j - \sum_i h_i S_i$$

# Example 1: Boltzmann Machine

★ Fixed point :

$$\nabla \mathcal{L}_{\theta} = 0$$

$$\left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\mathcal{D}}} = \left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\theta}} \quad \forall \theta_i$$

Moment matching statistics

$$\frac{\partial \mathcal{L}}{\partial J_{ij}} = \langle S_i S_j \rangle_{p_{\mathcal{D}}} - \langle S_i S_j \rangle_{p_{\theta}}$$

$$\frac{\partial \mathcal{L}}{\partial h_i} = \langle S_i \rangle_{p_{\mathcal{D}}} - \langle S_i \rangle_{p_{\theta}}$$

Ising-like model

$$E_{J, h}(S) = - \sum_{ij} J_{ij} S_i S_j - \sum_i h_i S_i$$

$$\frac{\partial E}{\partial J_{ij}} = -S_i S_j$$

$$\frac{\partial E}{\partial h_i} = -S_i$$

# Example 1: Boltzmann Machine

★ Fixed point :  $\nabla \mathcal{L}_{\theta} = 0$

$$\left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\mathcal{D}}} = \left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\theta}} \quad \forall \theta_i$$

Moment matching statistics

$$\frac{\partial \mathcal{L}}{\partial J_{ij}} = \langle S_i S_j \rangle_{p_{\mathcal{D}}} - \langle S_i S_j \rangle_{p_{\theta}}$$

$$\frac{\partial \mathcal{L}}{\partial h_i} = \langle S_i \rangle_{p_{\mathcal{D}}} - \langle S_i \rangle_{p_{\theta}}$$

Ising-like model

$$E_{J,h}(S) = - \sum_{ij} J_{ij} S_i S_j - \sum_i h_i S_i$$

$$\frac{\partial E}{\partial J_{ij}} = -S_i S_j \quad \frac{\partial E}{\partial h_i} = -S_i$$

We can encode the covariance matrix of the data but nothing beyond that!

Solution  
is unique !

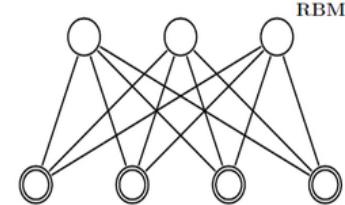
Fixed point

$$\langle S_i S_j \rangle_{p_{J,h}} = \langle S_i S_j \rangle_{p_{\mathcal{D}}}$$

$$\langle S_i \rangle_{p_{J,h}} = \langle S_i \rangle_{p_{\mathcal{D}}}$$

# Example 2: Restricted Boltzmann Machine

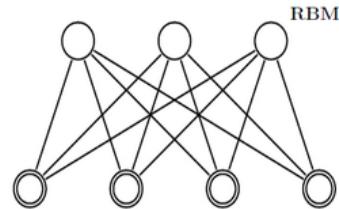
$$\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h}) = -\mathbf{x}^\top W \mathbf{h} - \zeta^\top \mathbf{x} - \eta^\top \mathbf{h} \quad p_{\theta}(\mathbf{x}) = \frac{\sum_{\mathbf{h}} e^{-\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h})}}{Z_{\theta}}$$



$$h_a = \{0, 1\} \quad \Rightarrow E_{\theta}(\mathbf{x}) = -\sum_i x_i \zeta_i - \sum_{a=1}^{N_h} \log \left( 1 + e^{\sum_i x_i W_{ia} + \eta_a} \right)$$

# Example 2: Restricted Boltzmann Machine

$$\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h}) = -\mathbf{x}^\top W \mathbf{h} - \zeta^\top \mathbf{x} - \eta^\top \mathbf{h} \quad p_{\theta}(\mathbf{x}) = \frac{\sum_{\mathbf{h}} e^{-\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h})}}{Z_{\theta}}$$



$$h_a = \{0, 1\} \Rightarrow E_{\theta}(\mathbf{x}) = -\sum_i x_i \zeta_i - \sum_{a=1}^{N_h} \log \left( 1 + e^{\sum_i x_i W_{ia} + \eta_a} \right)$$

$$\frac{\partial E}{\partial W_{ia}} = -\sigma \left( \sum_a W_{ia} \mathbf{x}_i + \eta_a \right) \mathbf{x}_i$$

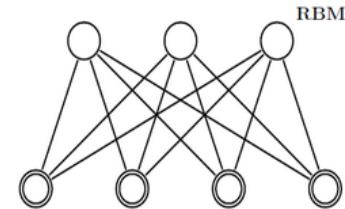
$$\frac{\partial E}{\partial \eta_a} = -\sigma \left( \sum_a W_{ia} \mathbf{x}_i + \eta_a \right)$$

$$\frac{\partial E}{\partial \zeta_i} = -\mathbf{x}_i$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} = \text{sigmoid}(x)$$

# Example 2: Restricted Boltzmann Machine

$$\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h}) = -\mathbf{x}^\top W \mathbf{h} - \zeta^\top \mathbf{x} - \eta^\top \mathbf{h} \quad p_{\theta}(\mathbf{x}) = \frac{\sum_{\mathbf{h}} e^{-\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h})}}{Z_{\theta}}$$



$$h_a = \{0, 1\} \Rightarrow E_{\theta}(\mathbf{x}) = -\sum_i x_i \zeta_i - \sum_{a=1}^{N_h} \log \left( 1 + e^{\sum_i x_i W_{ia} + \eta_a} \right)$$

$$\frac{\partial E}{\partial W_{ia}} = -\sigma \left( \sum_a W_{ia} \mathbf{x}_i + \eta_a \right) \mathbf{x}_i$$

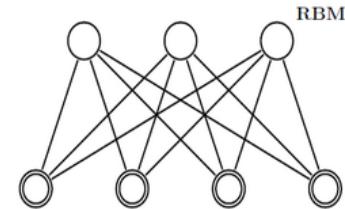
$$\frac{\partial E}{\partial \eta_a} = -\sigma \left( \sum_a W_{ia} \mathbf{x}_i + \eta_a \right)$$

$$\frac{\partial E}{\partial \zeta_i} = -\mathbf{x}_i$$

$$p(h_a = 1 | \mathbf{x}, \mathbf{h}_{-a}, \theta) = \frac{e^{\sum_i W_{ia} x_i + \eta_a}}{1 + e^{\sum_i W_{ia} x_i + \eta_a}} \\ = \sigma \left( \sum_i W_{ia} x_i + \eta_a \right) = \langle h_a \rangle_{p_{\mathcal{E}}(\mathbf{h} | \mathbf{x})}$$

# Example 2: Restricted Boltzmann Machine

$$\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h}) = -\mathbf{x}^\top W \mathbf{h} - \zeta^\top \mathbf{x} - \eta^\top \mathbf{h} \quad p_{\theta}(\mathbf{x}) = \frac{\sum_{\mathbf{h}} e^{-\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h})}}{Z_{\theta}}$$



$$h_a = \{0, 1\} \Rightarrow E_{\theta}(\mathbf{x}) = -\sum_i x_i \zeta_i - \sum_{a=1}^{N_h} \log \left( 1 + e^{\sum_i \mathbf{x}_i W_{ia} + \eta_a} \right)$$

$$\frac{\partial E}{\partial W_{ia}} = -\sigma \left( \sum_a W_{ia} \mathbf{x}_i + \eta_a \right) x_i = -\langle h_a \rangle_{p_{\mathcal{E}}(\mathbf{h} | \mathbf{x})} x_i$$

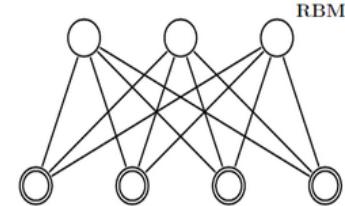
$$\frac{\partial E}{\partial \eta_a} = -\sigma \left( \sum_a W_{ia} \mathbf{x}_i + \eta_a \right) = -\langle h_a \rangle_{p_{\mathcal{E}}(\mathbf{h} | \mathbf{x})}$$

$$\frac{\partial E}{\partial \zeta_i} = -x_i$$

$$p(h_a = 1 | \mathbf{x}, \mathbf{h}_{-a}, \theta) = \frac{e^{\sum_i W_{ia} x_i + \eta_a}}{1 + e^{\sum_i W_{ia} x_i + \eta_a}} \\ = \sigma \left( \sum_i W_{ia} x_i + \eta_a \right) = \langle h_a \rangle_{p_{\mathcal{E}}(\mathbf{h} | \mathbf{x})}$$

# Example 2: Restricted Boltzmann Machine

$$\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{h}) = -\mathbf{x}^\top W \mathbf{h} - \boldsymbol{\zeta}^\top \mathbf{x} - \boldsymbol{\eta}^\top \mathbf{h} \quad p_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{\sum_{\mathbf{h}} e^{-\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{h})}}{Z_{\boldsymbol{\theta}}}$$



$$h_a = \{0, 1\} \quad \Rightarrow E_{\boldsymbol{\theta}}(\mathbf{x}) = -\sum_i x_i \zeta_i - \sum_{a=1}^{N_h} \log \left( 1 + e^{\sum_i x_i W_{ia} + \eta_a} \right)$$

$$p_{\mathcal{E}}(\mathbf{x}, \mathbf{h}) = p_{\mathcal{E}}(\mathbf{h} | \mathbf{x}) p_{\boldsymbol{\theta}}(\mathbf{x})$$

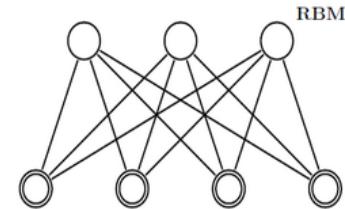
$$\frac{\partial \mathcal{L}}{\partial W_{ia}} = \left\langle x_i \langle h_a \rangle_{p_{\mathcal{E}}(\mathbf{h} | \mathbf{x})} \right\rangle_{p_{\mathcal{D}}} - \left\langle x_i \langle h_a \rangle_{p_{\mathcal{E}}(\mathbf{h} | \mathbf{x})} \right\rangle_{p_{\boldsymbol{\theta}}}$$

$$\frac{\partial \mathcal{L}}{\partial \eta_a} = \left\langle \langle h_a \rangle_{p_{\mathcal{E}}(\mathbf{h} | \mathbf{x})} \right\rangle_{p_{\mathcal{D}}} - \left\langle \langle h_a \rangle_{p_{\mathcal{E}}(\mathbf{h} | \mathbf{x})} \right\rangle_{p_{\boldsymbol{\theta}}}$$

$$\frac{\partial \mathcal{L}}{\partial \zeta_i} = \langle x_i \rangle_{p_{\mathcal{D}}} - \langle x_i \rangle_{p_{\boldsymbol{\theta}}}$$

# Example 2: Restricted Boltzmann Machine

$$\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h}) = -\mathbf{x}^\top W \mathbf{h} - \boldsymbol{\zeta}^\top \mathbf{x} - \boldsymbol{\eta}^\top \mathbf{h} \quad p_{\theta}(\mathbf{x}) = \frac{\sum_{\mathbf{h}} e^{-\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h})}}{Z_{\theta}}$$



$$h_a = \{0, 1\} \Rightarrow E_{\theta}(\mathbf{x}) = -\sum_i x_i \zeta_i - \sum_{a=1}^{N_h} \log \left( 1 + e^{\sum_i x_i W_{ia} + \eta_a} \right)$$

$$p_{\mathcal{E}}(\mathbf{x}, \mathbf{h}) = p_{\mathcal{E}}(\mathbf{h} | \mathbf{x}) p_{\theta}(\mathbf{x})$$

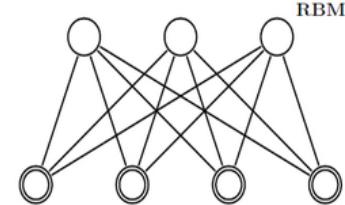
$$\frac{\partial \mathcal{L}}{\partial W_{ia}} = \left\langle x_i \langle h_a \rangle_{p_{\mathcal{E}}(\mathbf{h} | \mathbf{x})} \right\rangle_{p_{\mathcal{D}}} - \langle x_i h_a \rangle_{\mathcal{E}}$$

$$\frac{\partial \mathcal{L}}{\partial \eta_a} = \left\langle \langle h_a \rangle_{p_{\mathcal{E}}(\mathbf{h} | \mathbf{x})} \right\rangle_{p_{\mathcal{D}}} - \langle h_a \rangle_{\mathcal{E}}$$

$$\frac{\partial \mathcal{L}}{\partial \zeta_i} = \langle x_i \rangle_{p_{\mathcal{D}}} - \langle x_i \rangle_{\mathcal{E}}$$

# Example 2: Restricted Boltzmann Machine

$$\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h}) = -\mathbf{x}^\top W \mathbf{h} - \zeta^\top \mathbf{x} - \eta^\top \mathbf{h} \quad p_{\theta}(\mathbf{x}) = \frac{\sum_{\mathbf{h}} e^{-\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h})}}{Z_{\theta}}$$



$$h_a = \{0, 1\} \quad \Rightarrow E_{\theta}(\mathbf{x}) = -\sum_i x_i \zeta_i - \sum_{a=1}^{N_h} \log \left( 1 + e^{\sum_i x_i W_{ia} + \eta_a} \right)$$

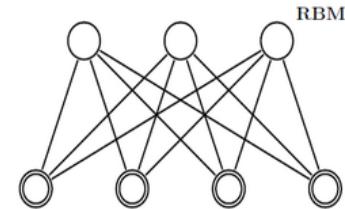
$$\frac{\partial \mathcal{L}}{\partial W_{ia}} = \langle x_i h_a \rangle_{p_{\mathcal{D}}} - \langle x_i h_a \rangle_{\mathcal{E}}$$

$$\frac{\partial \mathcal{L}}{\partial \eta_a} = \langle h_a \rangle_{p_{\mathcal{D}}} - \langle h_a \rangle_{\mathcal{E}}$$

$$\frac{\partial \mathcal{L}}{\partial \zeta_i} = \langle x_i \rangle_{p_{\mathcal{D}}} - \langle x_i \rangle_{\mathcal{E}}$$

# Example 2: Restricted Boltzmann Machine

$$\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h}) = -\mathbf{x}^\top W \mathbf{h} - \zeta^\top \mathbf{x} - \eta^\top \mathbf{h} \quad p_{\theta}(\mathbf{x}) = \frac{\sum_{\mathbf{h}} e^{-\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h})}}{Z_{\theta}}$$



$$h_a = \{0, 1\} \quad \Rightarrow E_{\theta}(\mathbf{x}) = -\sum_i x_i \zeta_i - \sum_{a=1}^{N_h} \log \left( 1 + e^{\sum_i x_i W_{ia} + \eta_a} \right)$$

Boltzmann machine:

$$\frac{\partial \mathcal{L}}{\partial J_{ij}} = \langle S_i S_j \rangle_{p_{\mathcal{D}}} - \langle S_i S_j \rangle_{p_{\theta}}$$

$$\frac{\partial \mathcal{L}}{\partial h_i} = \langle S_i \rangle_{p_{\mathcal{D}}} - \langle S_i \rangle_{p_{\theta}}$$

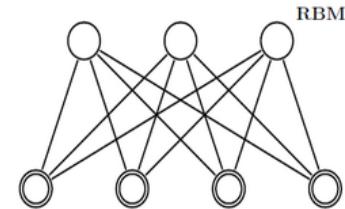
$$\frac{\partial \mathcal{L}}{\partial W_{ia}} = \langle x_i h_a \rangle_{p_{\mathcal{D}}} - \langle x_i h_a \rangle_{\mathcal{E}}$$

$$\frac{\partial \mathcal{L}}{\partial \eta_a} = \langle h_a \rangle_{p_{\mathcal{D}}} - \langle h_a \rangle_{\mathcal{E}}$$

$$\frac{\partial \mathcal{L}}{\partial \zeta_i} = \langle x_i \rangle_{p_{\mathcal{D}}} - \langle x_i \rangle_{\mathcal{E}}$$

# Example 2: Restricted Boltzmann Machine

$$\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h}) = -\mathbf{x}^\top W \mathbf{h} - \zeta^\top \mathbf{x} - \eta^\top \mathbf{h} \quad p_{\theta}(\mathbf{x}) = \frac{\sum_{\mathbf{h}} e^{-\mathcal{E}_{\theta}(\mathbf{x}, \mathbf{h})}}{Z_{\theta}}$$



$$h_a = \{0, 1\} \Rightarrow E_{\theta}(\mathbf{x}) = -\sum_i x_i \zeta_i - \sum_{a=1}^{N_h} \log \left( 1 + e^{\sum_i x_i W_{ia} + \eta_a} \right)$$

Boltzmann machine:

$$\frac{\partial \mathcal{L}}{\partial J_{ij}} = \langle S_i S_j \rangle_{p_{\mathcal{D}}} - \langle S_i S_j \rangle_{p_{\theta}}$$

$$\frac{\partial \mathcal{L}}{\partial h_i} = \langle S_i \rangle_{p_{\mathcal{D}}} - \langle S_i \rangle_{p_{\theta}}$$

$$\frac{\partial \mathcal{L}}{\partial W_{ia}} = \langle x_i h_a \rangle_{p_{\mathcal{D}}} - \langle x_i h_a \rangle_{\mathcal{E}}$$

$$\frac{\partial \mathcal{L}}{\partial \eta_a} = \langle h_a \rangle_{p_{\mathcal{D}}} - \langle h_a \rangle_{\mathcal{E}}$$

$$\frac{\partial \mathcal{L}}{\partial \zeta_i} = \langle x_i \rangle_{p_{\mathcal{D}}} - \langle x_i \rangle_{\mathcal{E}}$$



# Sample generation

# Generating new samples

<i>Empirical</i>	<i>Model</i>
$p_{\mathcal{D}}(\mathbf{x}) \sim \frac{e^{-E_{\boldsymbol{\theta}}(\mathbf{x})}}{Z_{\boldsymbol{\theta}}}$	Dominated minimum free-energy configurations
	$\{\mathbf{x}\}_{\text{eq}, \boldsymbol{\theta}} \sim \mathcal{D}$

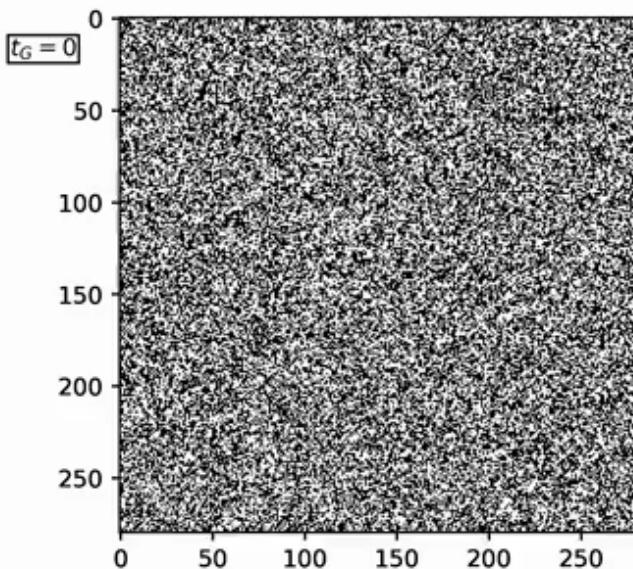
# Generating new samples

<i>Empirical</i>	<i>Model</i>
$p_{\mathcal{D}}(\mathbf{x}) \sim \frac{e^{-E_{\boldsymbol{\theta}}(\mathbf{x})}}{Z_{\boldsymbol{\theta}}}$	Dominated minimum free-energy configurations
	$\{\mathbf{x}\}_{\text{eq}, \boldsymbol{\theta}} \sim \mathcal{D}$
	$\left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\boldsymbol{\theta}}} \approx \left\langle \frac{\partial E}{\partial \theta_i} \right\rangle_{p_{\mathcal{D}}} \quad \forall \theta_i$

# Generating new samples

*Empirical*

$$p_{\mathcal{D}}(\mathbf{x}) \sim \frac{e^{-E_{\boldsymbol{\theta}}(\mathbf{x})}}{Z_{\boldsymbol{\theta}}}$$



*Model*

Dominated minimum  
free-energy  
configurations

$$\{\mathbf{x}\}_{\text{eq}, \boldsymbol{\theta}} \sim \mathcal{D}$$

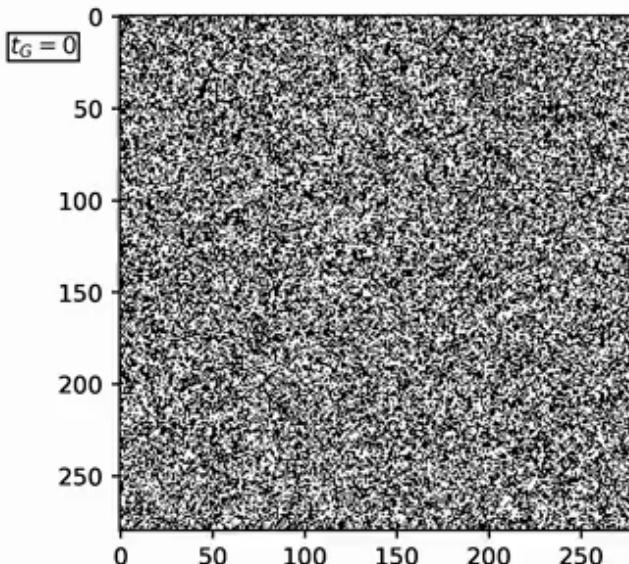


**Markov Chain  
Monte Carlo (MCMC)  
Langevin dynamics**

# Generating new samples

Empirical

$$p_{\mathcal{D}}(\mathbf{x}) \sim \frac{e^{-E_{\boldsymbol{\theta}}(\mathbf{x})}}{Z_{\boldsymbol{\theta}}}$$



Model

Dominated minimum  
free-energy  
configurations

$$\{\mathbf{x}\}_{\text{eq}, \boldsymbol{\theta}} \sim \mathcal{D}$$



Markov Chain  
Monte Carlo (MCMC)  
Langevin dynamics

$E_{\boldsymbol{\theta}}(\mathbf{x})$  Effective model  
for the data

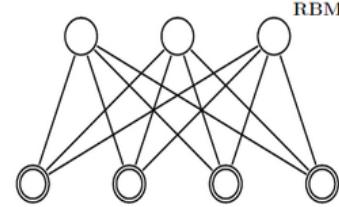
⇒ Free-energy landscape

If simple, we  
can analyze it!

Modeling, interpretability



# **Why Restricted Boltzmann Machines (RBMs) are good for that?**



# Why RBMs?

- **Simple enough** to allow some level of analytical treatment (MF)

- Phase diagram

A Decelle, C Furtlehner - Chinese Physics B, 2021

- Learning : sub-sequence of phase transitions

J Tubiana, R Monasson - Physical review letters, 2017

A Decelle, G Fissore, C Furtlehner - Journal of Statistical Physics, 2018

- Approximate methods to compute the free energy (TAP eqs.)

A Decelle, G Fissore, C Furtlehner  
Europhysics Letters, 2017

Biroli, Decelle, Bachtis, Seoane (2024, in prep.)

- Can be mapped to a physical interacting system

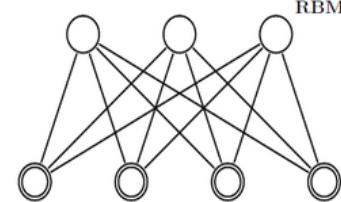
Gabrié, M., Tramel, E. W., & Krzakala, F. NeurIPS (2015)

Tramel, E. W., Gabrié, M., Manoel, A., Caltagirone, F., & Krzakala, F. Physical Review X (2018)

Decelle, A., Rosset, L., & Seoane, B. PRE (2023)

Decelle, Furtlehner, Navas & Seoane, B. SciPost Phys (2024)

- They are **expressive** : they can describe interesting datasets
- They are **frugal** models : fast code and to train
- They are **sample efficient** : perform well with small amounts of data



# Why Restricted Boltzmann Machines

- **Simple enough** to allow some level of analytical treatment (MF)

- Phase diagram
- Learning : sub-sequence of phase transitions

**Aurélien Decelle's  
lecture tomorrow**

- Approximate methods to compute the free energy (TAP eqs.)

Gabrié, M., Tramel, E. W., & Krzakala, F. NeurIPS (2015)

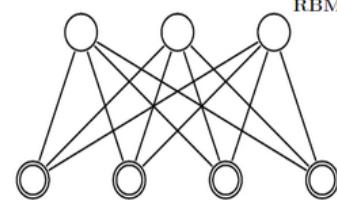
Tramel, E. W., Gabrié, M., Manoel, A., Caltagirone, F., & Krzakala, F. Physical Review X (2018)

- Can be mapped to a physical interacting system

Decelle, A., Rosset, L., & Seoane, B. PRE (2023)

Decelle, Furtlehner, Navas & Seoane, B. SciPost Phys (2024)

- They are **expressive** : they can describe interesting datasets
- They are **frugal** models : fast code and to train
- They are **sample efficient** : perform well with small amounts of data



# Why Restricted Boltzmann Machines

- Simple enough to allow some level of analytical treatment (MF)
  - Phase diagram
  - Learning : sub-sequence of phase transitions
  - Approximate methods to compute the free energy (TAP eqs.)

A Decelle, C Furtlehner - Chinese Physics B, 2021

J Tubiana, R Monasson - Physical review letters, 2017

A Decelle, G Fissore, C Furtlehner - Journal of Statistical Physics, 2018

A Decelle, G Fissore, C Furtlehner  
Europhysics Letters, 2017

Biroli, Decelle, Bachtis, Seoane (2024, in prep.)

- Can be mapped to a physical interacting system

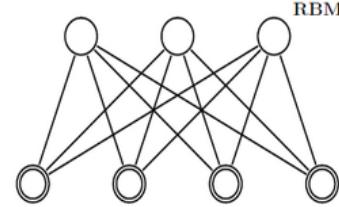
Gabrié, M., Tramel, E. W., & Krzakala, F. NeurIPS (2015)

Tramel, E. W., Gabrié, M., Manoel, A., Caltagirone, F., & Krzakala, F. Physical Review X (2018)

Decelle, A., Rosset, L., & Seoane, B. PRE (2023)

Decelle, Furtlehner, Navas & Seoane, B. SciPost Phys (2024)

- They are **expressive** : they can describe interesting datasets
- They are **frugal** models : fast code and to train
- They are **sample efficient** : perform well with small amounts of data



# Why Restricted Boltzmann Machines

## PLOS GENETICS

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

### Creating artificial human genomes using generative neural networks

Burak Yelmen, Aurélien Decelle, Linda Ongaro, Davide Marnetto, Corentin Tallec, Francesco Montinaro, Cyril Furtlehner, Luca Pagani, Flora Jay

SEARCH  
advanced search



## PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

### Deep convolutional and conditional neural networks for large-scale genomic data generation

Burak Yelmen, Aurélien Decelle, Leila Lea Boulos, Antoine Szatkowski, Cyril Furtlehner, Guillaume Charpiat, Flora Jay

Version 2 Published: October 30, 2023 • <https://doi.org/10.1371/journal.pcbi.1011584>

SEARCH  
advanced search



Currently the most accurate method to generate artificial human genome

Decelle, G Fissore, C Furtlehner, Journal of Statistical Physics, 2018

A Decelle, G Fissore, C Furtlehner, Europhysics Letters, 2017

Biroli, Decelle, Bachtis, Seoane (2024, in prep.)

free energy (TAP eqs.)

Gabrié, M., Tramel, E. W., & Krzakala, F. NeurIPS (2015)

Decelle, A., Caltagirone, F., & Krzakala, F. Physical Review X (2018)

Decelle, A., Rosset, L., & Seoane, B. PRE (2023)

system

Decelle, Furtlehner, Navas & Seoane, B. SciPost Phys (2024)

- They are **expressive** : they can describe interesting datasets
- They are **frugal** models : fast code and to train
- They are **sample efficient** : perform well with small amounts of data



# Why Restricted Boltzmann Machines



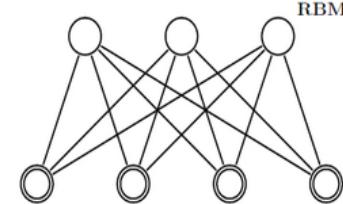
TOOLS AND RESOURCES



## Learning protein constitutive motifs from sequence data

Jérôme Tubiana, Simona Cocco, Rémi Monasson\*

Laboratory of Physics of the Ecole Normale Supérieure  
Research, Paris, France



They are able to capture biologically interpretable features related to function or structure...

PHYSICAL REVIEW LETTERS

Highlights Recent Accepted Collections Authors Referees Search Press About Editorial

### Mutational Paths with Sequence-Based Models of Proteins: From Sampling to Mean-Field Characterization

Eugenio Mauri, Simona Cocco, and Rémi Monasson  
Phys. Rev. Lett. **130**, 158402 – Published 12 April 2023

Article

References

Citing Articles (3)

Supplemental Material

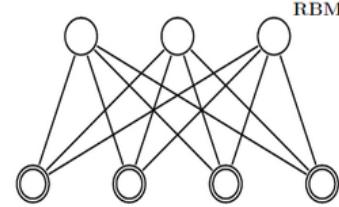
PDF

HTML

Export Citation

Propose mutational paths that can be validated in experiments

- They are **expressive** : they can describe interesting datasets
- They are **frugal** models : fast code and to train
- They are **sample efficient** : perform well with small amounts of data



# Why Restricted Boltzmann Machines

- Simple enough to allow some level of analytical treatment (MF)
  - If they are so cool, why are not they used more often?
    - Learning : sub-sequence of phase transitions  
→ EBMs are very difficult to train properly
    - Approximate methods to compute the free energy (TAP eqs.)
- They are expressive : they can describe interesting datasets
- They are frugal models : fast code and to train
- They are sample efficient : perform well with small amounts of data

A Decelle, C Furtlehner - Chinese Physics B, 2021

J Tubiana, R Monasson - Physical review letters, 2017

G Fissore, C Furtlehner - Chinese Physics B, 2018

A Decelle, G Fissore, C Furtlehner

Europhysics Letters, 2017

Biroli, Decelle, Bachtis, Seoane (2024, in prep.)

Gabrié, M., Tramel, E. W., & Krzakala, F. NeurIPS (2015)

Tramel, E. W., Gabrié, M., Manoel, A., Caltagirone, F., & Krzakala, F. Physical Review X (2018)

Decelle, A., Rosset, L., & Seoane, B. PRE (2023)

## Class 2 : Interpretability

- Can be mapped to a physical interacting system

## Class 3: Controlling the training

Decelle, Furtlehner, Navas & Seoane, B. SciPost Phys (2024)