Data modeling with Energy Based Models

Beatriz Seoane
LISN Paris-Saclay University

Acknowledgments

talento cm
Programa de atracción de talento investigador
Comunidad de Madrid

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

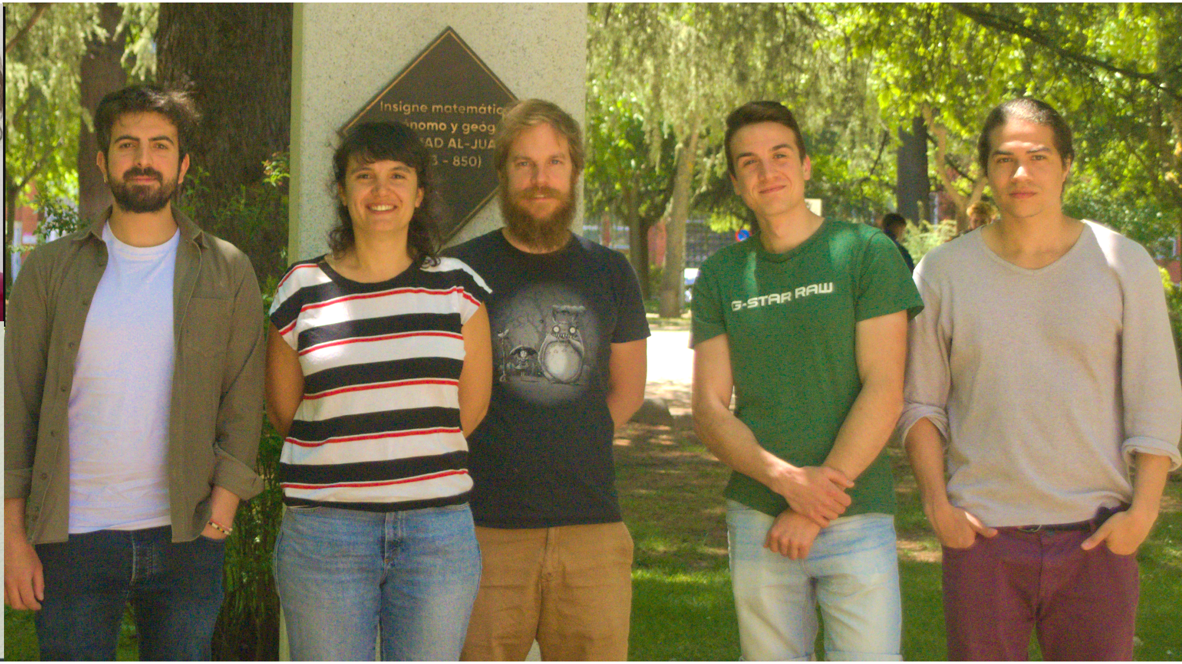
AGENCIA ESTATAL DE INVESTIGACIÓN

Santander

ANR
AGENCE NATIONALE DE LA RECHERCHE

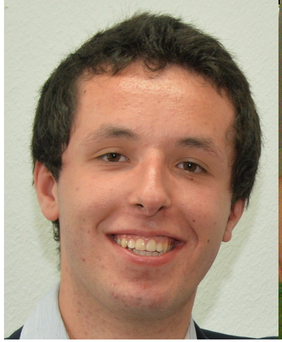
Aurélien Decelle
Giovanni Catania
Alfonso Navas
Lorenzo Rosset

UNIVERSIDAD COMPLUTENSE MADRID



Nicolas Béreux
Cyril Furtlehner

université PARIS-SACLAY



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

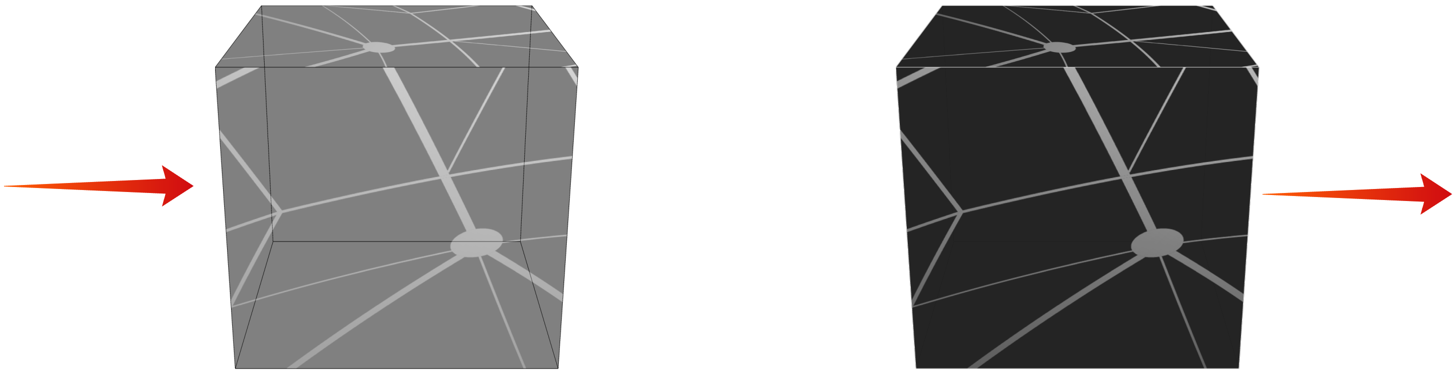
- Class 1: Introduction to Energy Based Models
    - Generative approach
    - Introduction to Energy-Based Models
        - The Restricted Boltzmann Machine (RBM)
    - Maximum likelihood training
    - Generation
    - Why I think RBMs are a cool tool

General definitions

Introduction : Generative approach

training

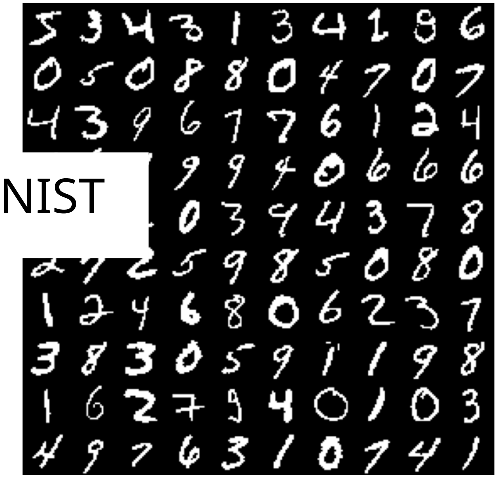
generating

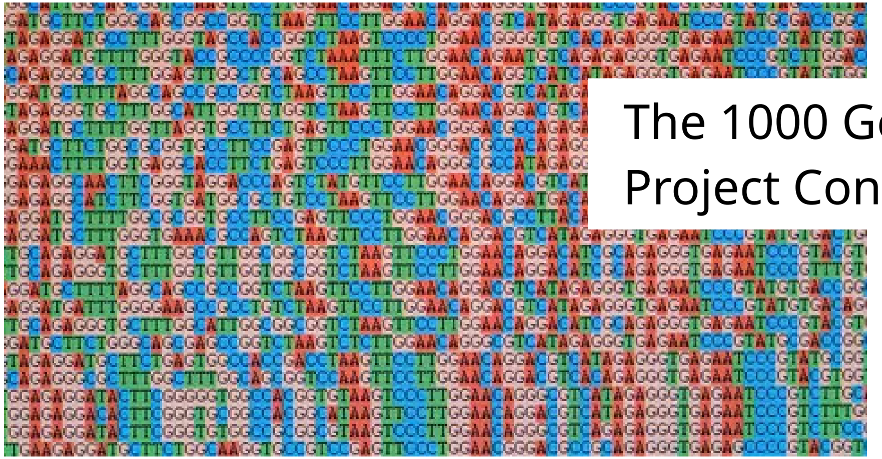


- **Energy based models (RBMs, Generative Convnets)**
- **Diffusion models,** normalizing flows, score based
- Variational AutoEncoder (VAE)
- Generative Adverarial Network (GAN)
- Autoregressive methods

Introduction : generative approach

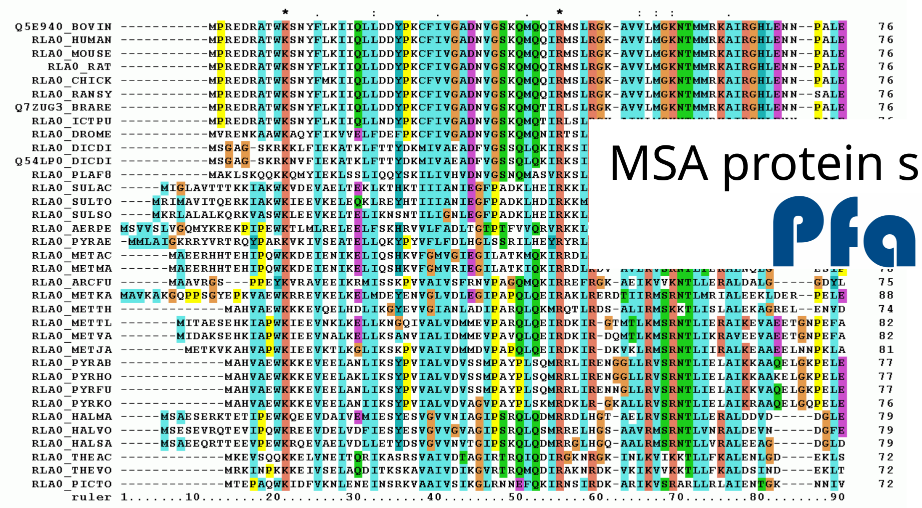
MNIST





The 1000 Genomes Project Consortium

CELEBA

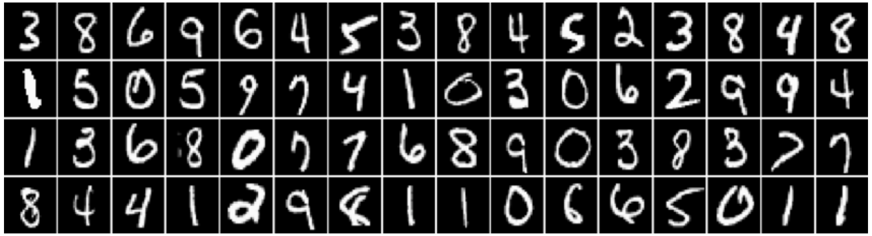


MSA protein sequences

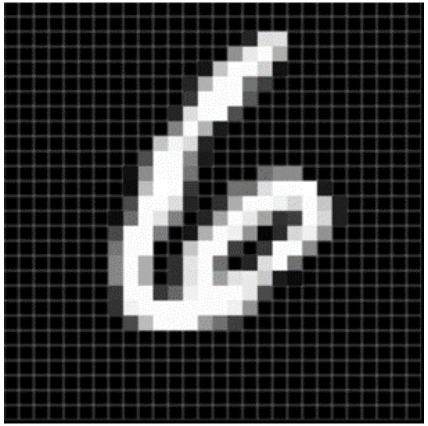
Data

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

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



M: # of examples in the data set



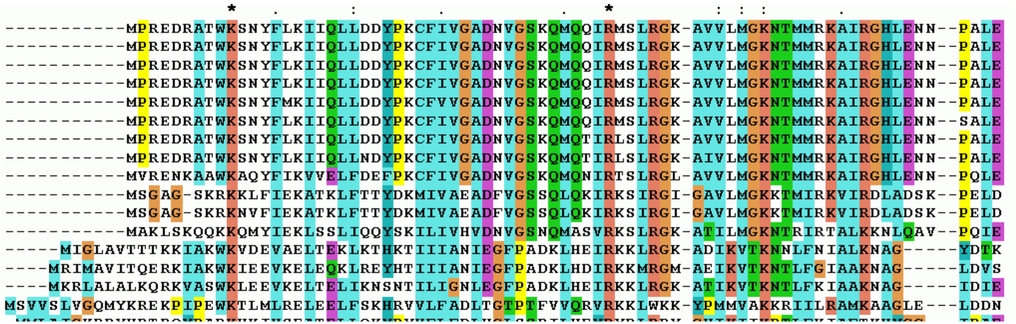
$$N = 28 \times 28$$

pixels

Data

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

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



M: # of sequences in a protein family

$N = L_{MSA}$ Amino-acids

Goal: Create synthetic sequences

Data

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

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

$\in [G, A \ldots, N, Q, -]^N$ categorical

Data distribution

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

Underlying assumption

i. i. d. realizations of a random variable

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

(Generally unknown)

Empirical data distribution

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

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

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

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

Empirical *Model*

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

Gibbs-Boltzmann distribution

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

$$Z_{\boldsymbol{\theta}} = \int d\boldsymbol{x}\, e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}$$

Partition function

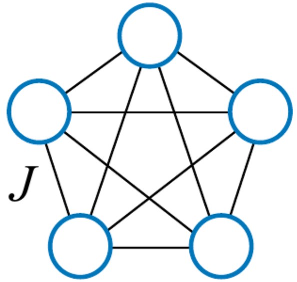
Learning : adjust the parameters $\boldsymbol{\theta}$ so that the dataset configurations are **typical** configurations of the model.

Energy based models (EBMs) Hinton, Hopfield, LeCun, Bengio

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

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.



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

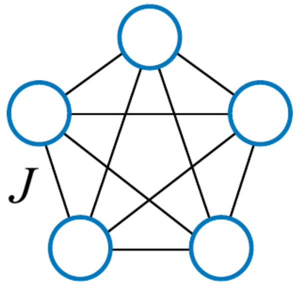
Pairwise interactions

Energy based models (EBMs)

Hinton, Hopfield, LeCun, Bengio

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

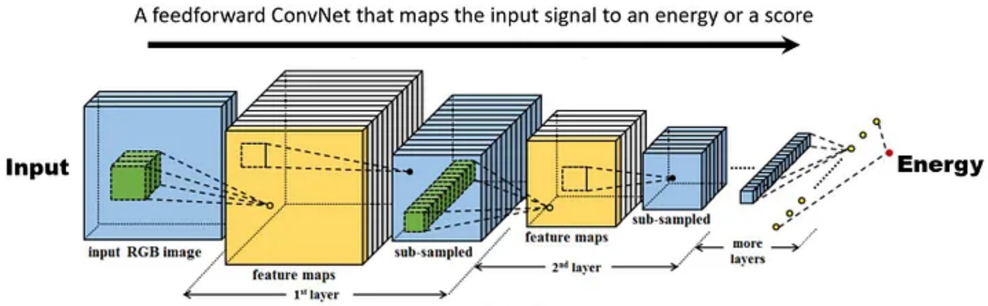
- Ising/Hopfield/Potts models

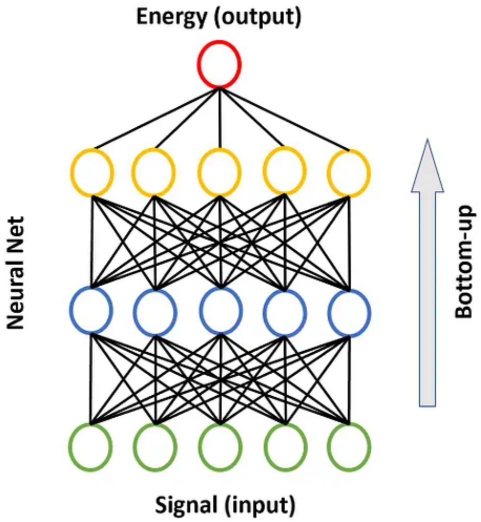


$$E_{J,\boldsymbol{h}}(\boldsymbol{x}) = -\boldsymbol{x}^\top J \boldsymbol{x} - \boldsymbol{h}^\top \boldsymbol{x}$$

- 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*.





Models with **hidden variables**

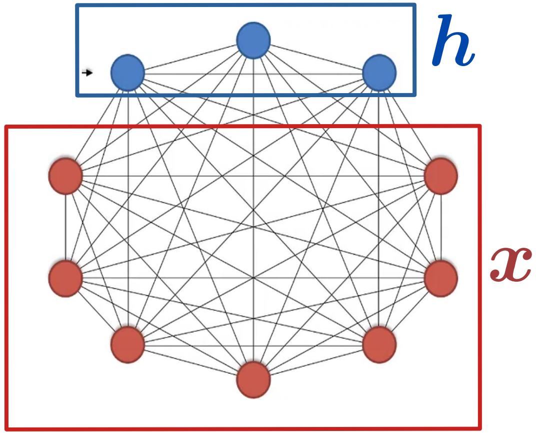
visible variables

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

Latent variables
Encode correlations

- Boltzmann Machines

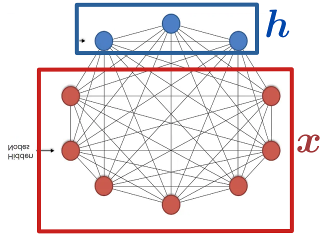
  - Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). *A learning algorithm for Boltzmann machines*. Cognitive science, 9(1), 147-169.



Models with **hidden variables**

visible variables

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



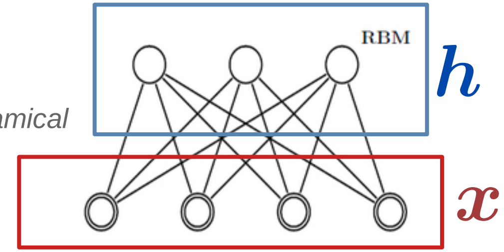
Latent variables
Encode correlations

- 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.*



Hidden layer : interactions

Visible layer : data

Models with hidden variables

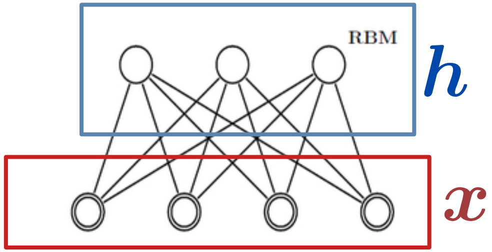
visible variables

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

Latent variables
Encode correlations

- Restricted Boltzmann Machine

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



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

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

Models with hidden variables

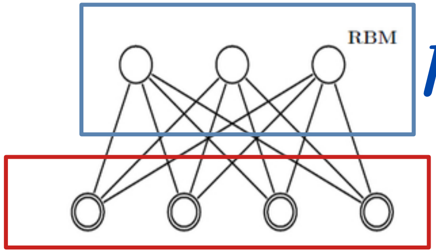
visible variables

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

Latent variables
Encode correlations

- Restricted Boltzmann Machine

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



$\boldsymbol{h}$

$\boldsymbol{x}$

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

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

$$= \frac{e^{\sum_i x_i \zeta_i}}{Z_{\boldsymbol{\theta}}} \prod_{a=1}^{N_{\mathrm{h}}} \left(1 + e^{\sum_i x_i W_{ia} + \eta_a}\right)$$

Models with hidden variables

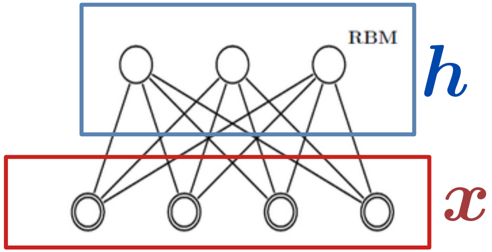
visible variables

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

Latent variables
Encode correlations

- Restricted Boltzmann Machine

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



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

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

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

$$= \frac{e^{\sum_i x_i \zeta_i}}{Z_{\boldsymbol{\theta}}} \prod_{a=1}^{N_{\mathrm{h}}} \left(1 + e^{\sum_i x_i W_{ia} + \eta_a}\right) \Rightarrow E_{\boldsymbol{\theta}}(\boldsymbol{x}) = -\sum_i x_i \zeta_i - \sum_{a=1}^{N_{\mathrm{h}}} \log\left(1 + e^{\sum_i x_i W_{ia} + \eta_a}\right)$$

Models with hidden variables

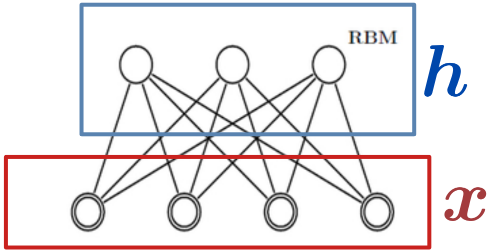
$$\mathcal{E}(\boldsymbol{x}, \boldsymbol{h}; \boldsymbol{\theta})$$

visible variables

Latent variables
Encode correlations

- Restricted Boltzmann Machine

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



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

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

$$= \frac{e^{\sum_i x_i \zeta_i}}{Z_{\boldsymbol{\theta}}} \prod_{a=1}^{N_\mathrm{h}} \left(1 + e^{\sum_i x_i W_{ia} + \eta_a}\right) \Rightarrow E_{\boldsymbol{\theta}}(\boldsymbol{x}) = -\sum_i x_i \zeta_i - \sum_{a=1}^{N_\mathrm{h}} \log\left(1 + e^{\sum_i x_i W_{ia} + \eta_a}\right)$$

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

Models with hidden variables

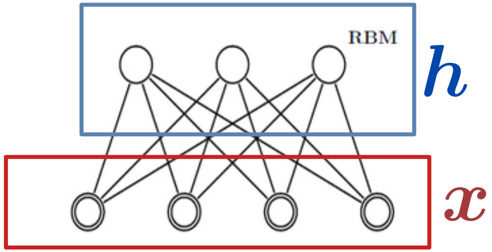
visible variables

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

Latent variables
Encode correlations

- Restricted Boltzmann Machine

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



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

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

$$= \frac{e^{\sum_i x_i \zeta_i}}{Z_{\boldsymbol{\theta}}} \prod_{a=1}^{N_\mathrm{h}} \left(1 + e^{\sum_i x_i W_{ia} + \eta_a}\right) \Rightarrow E_{\boldsymbol{\theta}}(\boldsymbol{x}) = -\sum_i x_i \zeta_i - \sum_{a=1} \log\left(1 + e^{\sum_i x_i W_{ia} + \eta_a}\right)$$

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

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

Le Roux and Bengio. Neural computation (2008)

Models with hidden variables

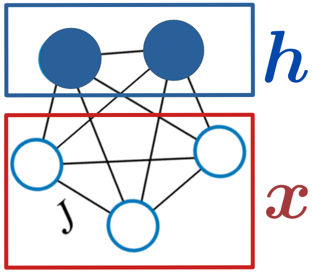
visible variables

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

Latent variables
Encode correlations

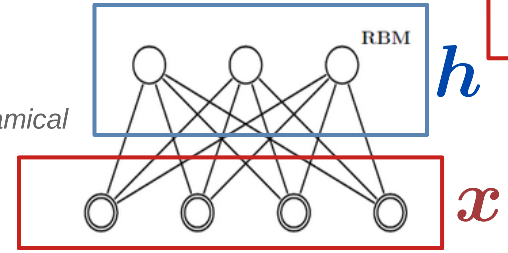
- Boltzmann Machines (Ising/Hopfield/Potts models)

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



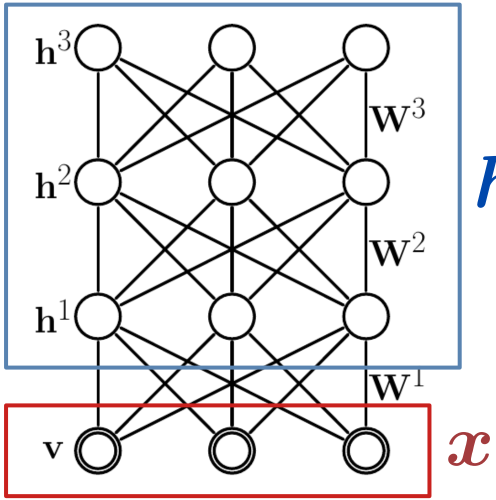
- Restricted Boltzmann Machine

  - 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.*

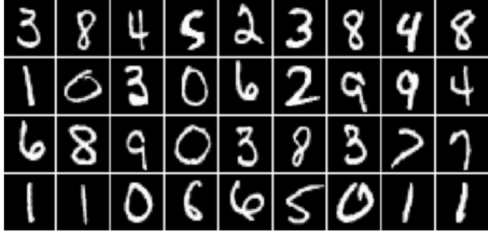


$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{h}} e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{h})}}{Z_{\boldsymbol{\theta}}}$$

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

Training procedure

Training procedure



Goal of the training: $\overset{\textit{Empirical}}{p_{\mathcal{D}}(\boldsymbol{x})} \sim \overset{\textit{Model}}{p_{\boldsymbol{\theta}}(\boldsymbol{x})} = \dfrac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$

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

Training procedure

Goal of the training:

$$\underbrace{p_{\mathcal{D}}(\boldsymbol{x})}_{\text{Empirical}} \sim \underbrace{p_{\boldsymbol{\theta}}(\boldsymbol{x})}_{\text{Model}} = \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

Minimize Kullback-Leibler (KL) divergence

$$D_{\mathrm{KL}}(p_{\mathcal{D}}||p_{\boldsymbol{\theta}}) = \int d\boldsymbol{x}\, p_{\mathcal{D}}(\boldsymbol{x}) \log \frac{p_{\mathcal{D}}(\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})}$$

$$= \underbrace{\int d\boldsymbol{x}\, p_{\mathcal{D}}(\boldsymbol{x}) \log p_{\mathcal{D}}(\boldsymbol{x})}_{\text{Constant}} - \int d\boldsymbol{x}\, p_{\mathcal{D}}(\boldsymbol{x}) \log p_{\boldsymbol{\theta}}(\boldsymbol{x})$$

Training procedure

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

Goal of the training:

$$\underbrace{p_{\mathcal{D}}(\boldsymbol{x})}_{\text{Empirical}} \sim \underbrace{p_{\boldsymbol{\theta}}(\boldsymbol{x})}_{\text{Model}} = \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

Minimize Kullback-Leibler (KL) divergence

$$D_{\mathrm{KL}}(p_{\mathcal{D}}||p_{\boldsymbol{\theta}}) = \int d\boldsymbol{x}\, p_{\mathcal{D}}(\boldsymbol{x}) \log \frac{p_{\mathcal{D}}(\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})}$$

$$= \underbrace{\int d\boldsymbol{x}\, p_{\mathcal{D}}(\boldsymbol{x}) \log p_{\mathcal{D}}(\boldsymbol{x})}_{\text{Constant}} - \int d\boldsymbol{x}\, p_{\mathcal{D}}(\boldsymbol{x}) \log p_{\boldsymbol{\theta}}(\boldsymbol{x})$$

log-likelihood

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

Training procedure

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

Empirical *Model*

Goal of the training: $$p_{\mathcal{D}}(\boldsymbol{x}) \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

Minimize Kullback-Leibler divergence $$D_{\mathrm{KL}}(p_{\mathcal{D}}||p_{\boldsymbol{\theta}}) \Longleftrightarrow$$ **Maximize** The log-likelihood $$\log L(\mathcal{D}|\boldsymbol{\theta}) \equiv \mathcal{L}(\mathcal{D}|\boldsymbol{\theta})$$

divergence

$$= \underbrace{\int d\boldsymbol{x}\, p_{\mathcal{D}}(\boldsymbol{x}) \log p_{\mathcal{D}}(\boldsymbol{x})}_{\text{Constant}} - \int d\boldsymbol{x}\, p_{\mathcal{D}}(\boldsymbol{x}) \log p_{\boldsymbol{\theta}}(\boldsymbol{x})$$

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

Training procedure

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

Empirical *Model*

Goal of the training: $$p_{\mathcal{D}}(\boldsymbol{x}) \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

Minimize Kullback-Leibler divergence $D_{\mathrm{KL}}(p_{\mathcal{D}}||p_{\boldsymbol{\theta}}) \Longleftrightarrow$ **Maximize** The log-likelihood $\log L(\mathcal{D}|\boldsymbol{\theta}) \equiv \mathcal{L}(\mathcal{D}|\boldsymbol{\theta})$

N

K

divergence

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

onstant

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

Recall Bayes-Theorem

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

Log-likelihood maximization

$$\boxed{\mathcal{L}(\mathcal{D}|\boldsymbol{\theta}) = \sum_{m=1}^{M} \log p_{\boldsymbol{\theta}} \left(\boldsymbol{x} = \boldsymbol{x}^{(m)} \right)} \qquad p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

Log-likelihood maximization

$$\boxed{\mathcal{L}(\mathcal{D}|\boldsymbol{\theta}) = \sum_{m=1}^{M} \log p_{\boldsymbol{\theta}}\left(\boldsymbol{x} = \boldsymbol{x}^{(m)}\right)} \qquad p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

Partition function

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

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

If x_i binary $\rightarrow 2^N$

Intractable

Log-likelihood maximization

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

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

Partition function

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

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

If x_i binary → 2^N

Intractable

(Stochastic) gradient **ascent**

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} \mathcal{L}$$

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

Log-likelihood maximization

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

Partition function

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

(Stochastic) gradient **ascent**

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} \mathcal{L}$$

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

Log-likelihood maximization

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

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

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

(Stochastic) gradient **ascent**

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} \mathcal{L}$$

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

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

Log-likelihood maximization

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

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

$$\overline{\boldsymbol{\nabla} E_{\boldsymbol{\theta}}}^{\mathcal{D}}$$

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

(Stochastic) gradient **ascent**

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} \mathcal{L}$$

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

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

Easy !

Log-likelihood maximization

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

$$\frac{\partial \log Z}{\partial \theta_i} = \sum_{\{\boldsymbol{x}\}} \frac{e^{-E(\boldsymbol{x})}}{Z} \frac{\partial E(\boldsymbol{x})}{\partial \theta_i} = \left\langle \frac{\partial E(\boldsymbol{x})}{\partial \theta_i} \right\rangle$$

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

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

MCMC sampling

(Stochastic) gradient **ascent**

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} \mathcal{L}$$

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

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

Easy !

Hard !

Log-likelihood

Every time we want to update the parameters

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

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

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

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

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

MCMC sampling

(Stochastic) gradient **ascent**

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} \mathcal{L}$$

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

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

Easy ! **Hard !**

Log-likelihood

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

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

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

<u>Every time</u> we want to update the parameters

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

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

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

(Stochastic) gradient **ascent**

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

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

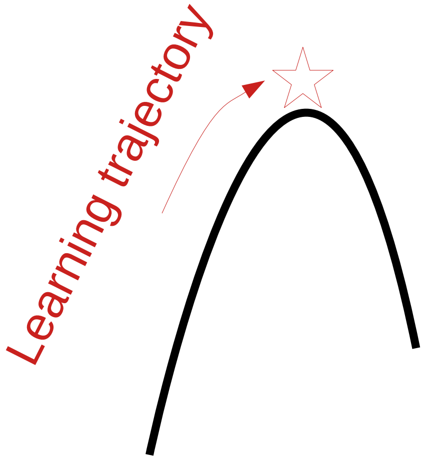
MCMC sampling

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

Easy !

Hard !

On the gradient ascent



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

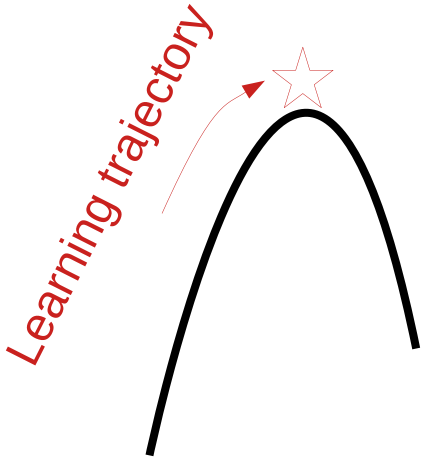
Update rule: $\boldsymbol{\nabla} \mathcal{L}_{\boldsymbol{\theta}} = \langle -\boldsymbol{\nabla} E_{\boldsymbol{\theta}} \rangle_{p_{\mathcal{D}}} - \langle -\boldsymbol{\nabla} E_{\boldsymbol{\theta}} \rangle_{p_{\boldsymbol{\theta}}}$

On the gradient ascent

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

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

Moment matching statistics



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

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

On the gradient ascent

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

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

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

Moment matching statistics

Hessian matrix

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

$$- \langle f_{\theta_i}(\boldsymbol{x}, \boldsymbol{\theta}) f_{\theta_j}(\boldsymbol{x}, \boldsymbol{\theta}) \rangle_{p_{\boldsymbol{\theta}}} + \langle f_{\theta_i}(\boldsymbol{x}, \boldsymbol{\theta}) \rangle_{p_{\boldsymbol{\theta}}} \langle f_{\theta_j}(\boldsymbol{x}, \boldsymbol{\theta}) \rangle_{p_{\boldsymbol{\theta}}}$$

Update rule:

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

On the gradient ascent

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

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

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

Moment matching statistics

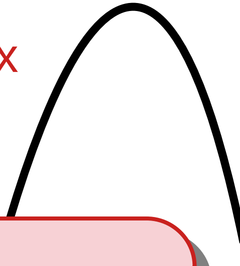
Hessian matrix

if

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

$$- \langle f_{\theta_i}(\boldsymbol{x}, \boldsymbol{\theta}) f_{\theta_j}(\boldsymbol{x}, \boldsymbol{\theta}) \rangle_{p_{\boldsymbol{\theta}}} + \langle f_{\theta_i}(\boldsymbol{x}, \boldsymbol{\theta}) \rangle_{p_{\boldsymbol{\theta}}} \langle f_{\theta_j}(\boldsymbol{x}, \boldsymbol{\theta}) \rangle_{p_{\boldsymbol{\theta}}}$$

Semi negative definite ⇒ convex



Update rule:

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

Example 1: Boltzmann Machine

☆ **Fixed point :** $\nabla \mathcal{L}_{\boldsymbol{\theta}} = \mathbf{0}$

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

Moment matching statistics

Ising-like model

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

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

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

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

Example 1: Boltzmann Machine

☆ **Fixed point :** $\boldsymbol{\nabla} \mathcal{L}_{\boldsymbol{\theta}} = \mathbf{0}$

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

Moment matching statistics

Ising-like model

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

Example 1: Boltzmann Machine

☆ **Fixed point :** $\boldsymbol{\nabla} \mathcal{L}_{\boldsymbol{\theta}} = \mathbf{0}$

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

Moment matching statistics

Ising-like model

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

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

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

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

Example 1: Boltzmann Machine

☆ **Fixed point :** $\boldsymbol{\nabla} \mathcal{L}_{\boldsymbol{\theta}} = \mathbf{0}$

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

Moment matching statistics

Ising-like model

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

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

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

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

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

Solution is unique !

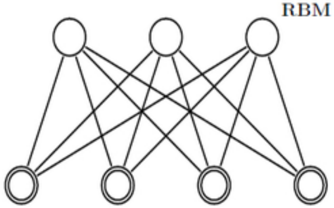
Fixed point

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

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

Example 2: Restricted Boltzmann Machine

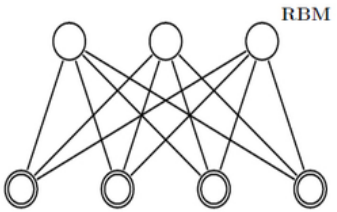
$$\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{h}) = -\boldsymbol{x}^\top W \boldsymbol{h} - \boldsymbol{\zeta}^\top \boldsymbol{x} - \boldsymbol{\eta}^\top \boldsymbol{h} \qquad p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{h}} e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{h})}}{Z_{\boldsymbol{\theta}}}$$



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

Example 2: Restricted Boltzmann Machine

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



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

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

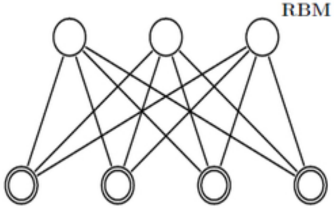
$$\frac{\partial E}{\partial \eta_a} = -\sigma\left(\sum_a W_{ia}x_i+\eta_a\right)$$

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

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

Example 2: Restricted Boltzmann Machine

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



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

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

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

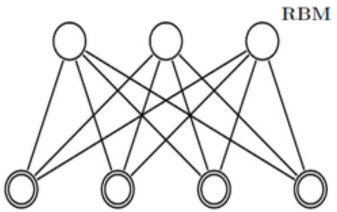
$$\frac{\partial E}{\partial \zeta_i} = -x_i$$

$$p(h_a = 1 | \boldsymbol{x}, \boldsymbol{h}_{-a}, \boldsymbol{\theta}) = \frac{e^{\sum_i W_{ia} x_i + \eta_a}}{1 + e^{\sum_i W_{ia} x_i + \eta_a}}$$

$$= \sigma\left(\sum_i W_{ia} x_i + \eta_a\right) = \langle h_a \rangle_{p_{\mathcal{E}}(\boldsymbol{h}|x)}$$

Example 2: Restricted Boltzmann Machine

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



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

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

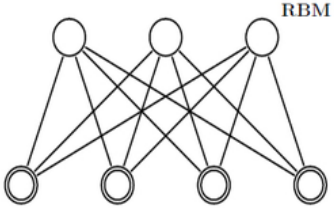
$$\frac{\partial E}{\partial \eta_a} = -\sigma\left(\sum_a W_{ia}x_i+\eta_a\right) = -\langle h_a\rangle_{p_{\mathcal{E}}(\boldsymbol{h}|x)}$$

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

$$p(h_a = 1|\boldsymbol{x},\boldsymbol{h}_{-a},\boldsymbol{\theta}) = \frac{e^{\sum_i W_{ia}x_i+\eta_a}}{1+e^{\sum_i W_{ia}x_i+\eta_a}}$$

$$= \sigma\left(\sum_i W_{ia}x_i + \eta_a\right) = \langle h_a\rangle_{p_{\mathcal{E}}(\boldsymbol{h}|x)}$$

Example 2: Restricted Boltzmann Machine



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

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

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

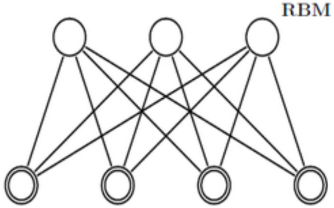
$$\frac{\partial \mathcal{L}}{\partial W_{ia}} = \left\langle x_i \left\langle h_a \right\rangle_{p_{\mathcal{E}}(\boldsymbol{h}|x)} \right\rangle_{p_{\mathcal{D}}} - \left\langle x_i \left\langle h_a \right\rangle_{p_{\mathcal{E}}(\boldsymbol{h}|x)} \right\rangle_{p_{\boldsymbol{\theta}}}$$

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

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

Example 2: Restricted Boltzmann Machine

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



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

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

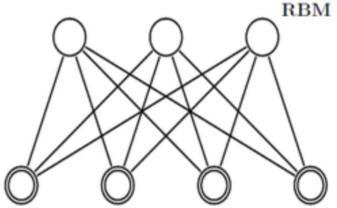
$$\frac{\partial\mathcal{L}}{\partial W_{ia}} = \left\langle x_i \left\langle h_a\right\rangle_{p_{\mathcal{E}}(\boldsymbol{h}|x)}\right\rangle_{p_{\mathcal{D}}} - \left\langle x_i h_a\right\rangle_{\mathcal{E}}$$

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

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

Example 2: Restricted Boltzmann Machine

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



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

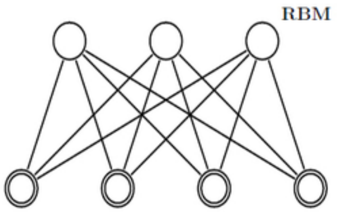
$$\frac{\partial\mathcal{L}}{\partial W_{ia}} = \langle x_i h_a\rangle_{p_\mathcal{D}} - \langle x_i h_a\rangle_{\mathcal{E}}$$

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

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

Example 2: Restricted Boltzmann Machine

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



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

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

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

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

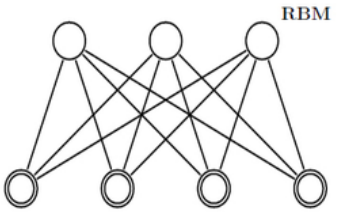
Boltzmann machine:

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

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

Example 2: Restricted Boltzmann Machine

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



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

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

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

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

Boltzmann machine:

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

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

Sample generation

Generating new samples

Empirical *Model*

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

Dominated minimum
free-energy
configurations

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

Generating new samples

Empirical *Model*

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

Dominated minimum free-energy configurations

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

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

Generating new samples

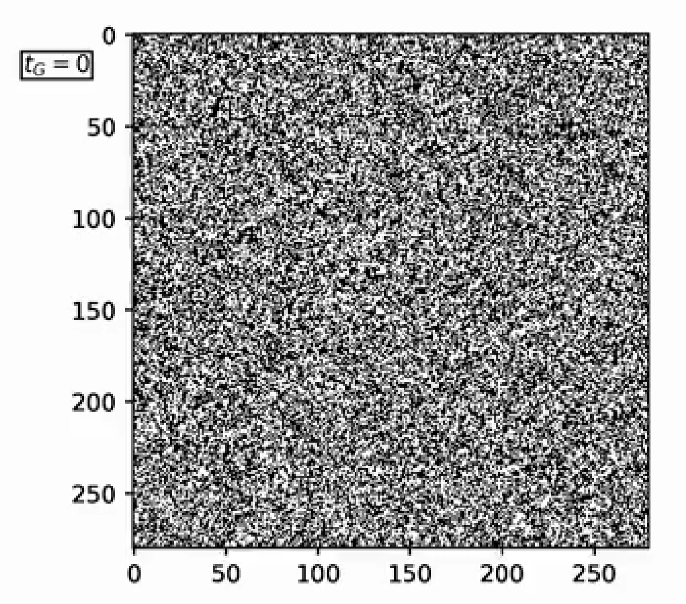
Empirical *Model*

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

Dominated minimum free-energy configurations

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

$\Longrightarrow$ **Markov Chain Monte Carlo (MCMC)**
Langevin dynamics



Generating new samples

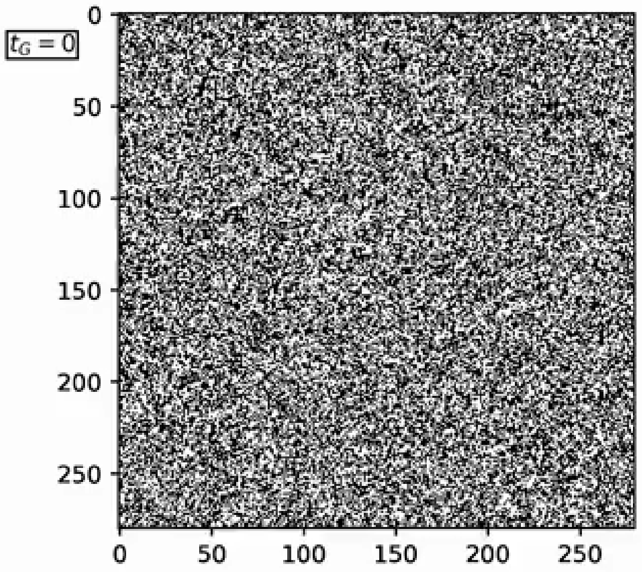
Empirical *Model*

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

Dominated minimum free-energy configurations

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

$\Longrightarrow$ **Markov Chain Monte Carlo (MCMC)**
Langevin dynamics



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

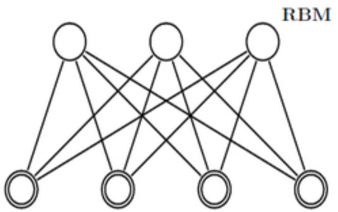
If simple, we can analyze it!

$\Rightarrow$ *Free-energy landscape*

Modeling, <u>interpretability</u>

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

Why RBMs?



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

A Decelle, C Furtlehner - Chinese Physics B, 2021

  - Phase diagram

J Tubiana, R Monasson - Physical review letters, 2017
A Decelle, G Fissore, C Furtlehner - Journal of Statistical Physics, 2018

  - Learning : sub-sequence of phase transitions

A Decelle, G Fissore, C Furtlehner
Europhysics Letters, 2017
Biroli, Decelle, Bachtis, Seoane (2024, in prep.)

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

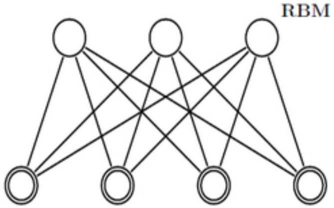
Gabrié, M., Tramel, E. W., & Krzakala, F. NeurIPS (2015)
Tramel, E. W., Gabrié, M., Manoel, A., Caltagirone, F., & Krzakala, F. Physical Review X (2018)
Decelle, A., Rosset, L., & Seoane, B. PRE (2023)

  - Can be mapped to a physical interacting system

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

- They are **expressive** : they can describe interesting datsets
- The 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)

  A Decelle, C Furtlehner - Chinese Physics B, 2021

  - Phase diagram

    Aurélien Decelle's lecture tomorrow

    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

  - Learning : sub-sequence of phase transitions

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

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

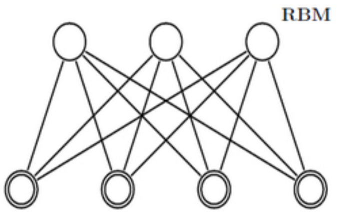
    Gabrié, M., Tramel, E. W., & Krzakala, F. NeurIPS (2015)
    Tramel, E. W., Gabrié, M., Manoel, A., Caltagirone, F., & Krzakala, F. Physical Review X (2018)
    Decelle, A., Rosset, L., & Seoane, B. PRE (2023)

  - Can be mapped to a physical interacting system

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

- They are **expressive** : they can describe interesting datsets
- The 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)

A Decelle, C Furtlehner - Chinese Physics B, 2021

  - Phase diagram

J Tubiana, R Monasson - Physical review letters, 2017
A Decelle, G Fissore, C Furtlehner - Journal of Statistical Physics, 2018

  - Learning : sub-sequence of phase transitions

A Decelle, G Fissore, C Furtlehner
Europhysics Letters, 2017
Biroli, Decelle, Bachtis, Seoane (2024, in prep.)

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

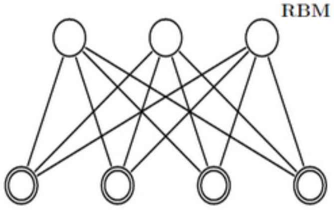
Gabrié, M., Tramel, E. W., & Krzakala, F. NeurIPS (2015)
Tramel, E. W., Gabrié, M., Manoel, A., Caltagirone, F., & Krzakala, F. Physical Review X (2018)
Decelle, A., Rosset, L., & Seoane, B. PRE (2023)

  - Can be mapped to a physical interacting system

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

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

Why Restricted Boltzmann Machines



BROWSE PUBLISH ABOUT SEARCH

advanced search

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

109 Save	78 Citation
20,493 View	102 Share

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 Szatkownik, Cyril Furtlehner, Guillaume Charpiat, Flora Jay

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

12 Save	2 Citation
1,520 View	15 Share

Currently the most accurate method to generate artificial human genome

ytical treatment (MF)

hysics B, 2021

w letters, 2017

Decelle, C Furtlehner, Journal of Statistical Physics, 2018

A Decelle, G Fissore, C Furtlehner

Europhysics Letters, 2017

itions

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

ee energy (TAP eqs.)

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

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

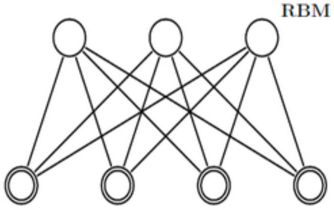
Decelle, A., Rosset, L., & Seoane, B. PRE (2023)

system

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

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

Why Restricted Boltzmann Machines



eLIFE
elifesciences.org

TOOLS AND RESOURCES

Learning protein constitutive motifs from sequence data

Jérôme Tubiana, Simona Cocco, Rémi Monasson*

Laboratory of Physics of the Ecole Normale Supérieur
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

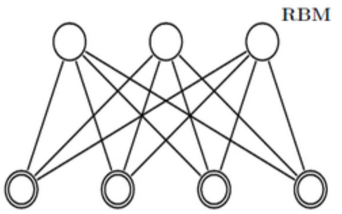
Eugenio Mauri, Simona Cocco, and Rémi Monasson
Phys. Rev. Lett. **130**, 158402 – Published 12 April 2023

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
- The are **frugal** models : fast code and to train
- They are **sample efficient** : perform well with small amounts of data

Why Restricted Boltzmann Machines



If they are so cool, why are not they used more often?

→ EBMs are very difficult to train properly

Class 2 : Interpretability

Class 3: Controlling the training

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

  A Decelle, C Furtlehner - Chinese Physics B, 2021

  J Tubiana, R Monasson - Physical review letters, 2017

  – Phase diagram

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

  – Learning : sub-sequence of phase transitions

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

  – 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)

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

  – Can be mapped to a physical interacting system

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

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