# Restricted Boltzmann Machines

## Use of non-linear-type layers and compositional phase in RBMs

Gianluca Manzan

gianluca.manzan@studenti.unipd.it

bioML Journal Club
University of Padova

May 19, 2021

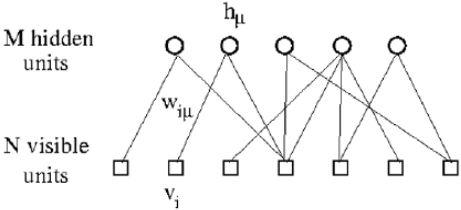
# Contents

1. RBM Introduction
2. Hopfield network
3. ReLU-RBM
4. Replica theory
5. Homogeneus free energy
6. Compositional phase

# RBM Introduction

- RBMs are unsupervised algoritms: learning features from data without any a priori knowledge
- RBM structure:

$$E[\boldsymbol{v},\boldsymbol{h}] = -\sum_{i=1}^{N}\sum_{\mu=1}^{M} w_{i\mu} v_i h_\mu - \sum_{i=1}^{N}\mathcal{U}_i(v_i) + \sum_{\mu=1}^{M}\mathcal{U}_\mu(h_\mu) \qquad (1)$$



J.Tubiana, R.Monasson (2017)

# RBM Introduction

- Change hidden units potential ($\mathcal{U}_\mu$) affects performance
- The dynamics of the network (Metropolis) gives Boltzmann distribution at equilibrium: $P(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z} e^{-E(\boldsymbol{v},\boldsymbol{h})/T}$
- Because of "Conditional independence property": $P(\boldsymbol{h}|\boldsymbol{v}) = \prod_\mu P(h_\mu|\boldsymbol{v})$ we can get the probability of a given unit to be active:

$$P(h_\mu|\boldsymbol{v}) = \frac{e^{h_\mu I_\mu - \mathcal{U}_\mu(h_\mu)}}{Z(\boldsymbol{h}|\boldsymbol{v})}$$

  where $I_\mu = \sum_i w_{i\mu} v_i$

Definition

$\phi_\mu :=$ activation function for $h_\mu$: most probable value $h^*_\mu$ of the hidden unit $\mu$ given the configuration $\boldsymbol{v}$

# RBM Introduction

- The most probable value is linked to the Energy-based behaviour of the RBM

$$h_\mu^* = \phi_\mu(I_\mu) = \operatorname*{argmax}_{h_\mu} P(h_\mu|\boldsymbol{v})$$

$$\phi_\mu(I_\mu) = \operatorname*{argmax}_{h_\mu} \frac{P(h_\mu, \boldsymbol{v})}{P(\boldsymbol{v})} = \operatorname*{argmax}_{h_\mu} \frac{e^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z}$$

$$\equiv \operatorname*{argmin}_{h_\mu} E(\boldsymbol{h}, \boldsymbol{v})$$

- Most probable hidden unit state $h_\mu^*$ is given by minimization of the Energy

# RBM Introduction

Most probable hidden unit state $h^*_\mu$ is given by minimization of the Energy

$$\frac{\partial E(\boldsymbol{h},\boldsymbol{v})}{\partial h_\mu} = \sum_{i,\nu} \left[ -w_{i\mu} v_i + \frac{\partial \mathcal{U}_\nu}{\partial h_\mu} \right] \delta_{\mu,\nu} \Bigg|_{h^*_\mu} = 0$$

$$-w_{i\mu} v_i + \frac{\partial \mathcal{U}_\mu}{\partial h_\mu}\left(h^*_\mu\right) = 0$$

$$\implies h^*_\mu = \left(\frac{\partial \mathcal{U}_\mu}{\partial h_\mu}\right)^{-1} (I_\mu)$$

# RBM Introduction

Consequences in the Gibbs sampling procedure of the gradient descent method:

1. Starting from initial $\boldsymbol{v}$
2. $I_\mu \rightarrow P(h_\mu|\boldsymbol{v}) \quad \forall_{\mu=1\dots M} \rightarrow \boldsymbol{h^*}$
3. $I_i \rightarrow P(v_i|\boldsymbol{h}) \quad \forall_{i=1\dots N} \rightarrow \boldsymbol{v}^*$

Learning algorithm:

$$\frac{\partial \mathcal{L}}{\partial w_{i,\mu}} = <v_i h_\mu>_{data} - <v_i h_\mu>_{model}$$

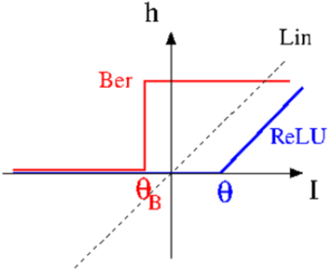
where $\mathcal{L}$ is the log likelihood

# RBM Introduction

Different activation functions:

| Property | Bernoulli | Gaussian | Rectified Linear |
|---|---|---|---|
| Domain | $\mathcal{X} \in \{0,1\}$ | $\mathcal{X} \in \mathbb{R}$ | $\mathcal{X} \in [0,+\infty)$ |
| Potential | $\mathcal{U} = -g\mathcal{X}$ | $\mathcal{U} = \frac{1}{2}\mathcal{X}^2$ | $\mathcal{U} = \frac{1}{2}\mathcal{X}^2 + \theta\mathcal{X}$ |
| Activation function | $\Phi = \Theta(x-g)$ | $\Phi = x$ | $\Phi = \max\{0, x-\theta\}$ |

Table 1: Activation functions

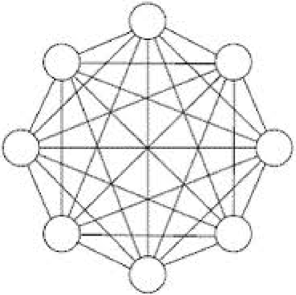


J.Tubiana, R.Monasson (2017)

# Hopfield model

The RBMs framework can describe many different kind of interactions between visible units, using its interanl representation ($\boldsymbol{h}$). An example is the Hopfield network.

$$E[\boldsymbol{v}] = -\sum_{i=1}\sum_{j=1} v_i J_{ij} v_j - \sum_{i=1} a_i(v_i)$$



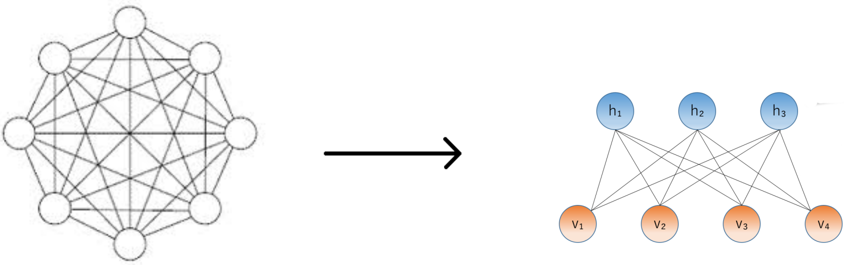
Hopfield network

# Hopfield network

The Hubbard-Stratonovich (HS) transformation can be performed to link the Hopfield network to its RBM counterpart, rewriting $J_{ij} = \sum_\mu w_{i\mu} w_{j\mu}$:

$$\mathsf{P}(\boldsymbol{v}) = \frac{1}{Z} e^{\sum_{ij\mu} v_i w_{i\mu} w_{j\mu} v_j + \sum_{i=1}^{N} a_i(v_i)} \to \frac{1}{Z} e^{\sum_{i=1}^{N} a_i(v_i)} \prod_\mu \int_\mu e^{\sum_\mu h_\mu^2 + \sum_{i\mu} v_i w_{i\mu} h_\mu}$$

$$E[\boldsymbol{v}] = -\sum_{i=1} \sum_j v_i J_{ij} v_j - \sum_i a_i(v_i) \to E[\boldsymbol{v}, \boldsymbol{h}] = -\sum_i a_i(v_i) + \frac{1}{2} \sum_\mu h_\mu^2 + \sum_{i\mu} w_{i\mu} v_i h_\mu$$



Hopfield network

RBM network

# Hopfield network

The above argument can be reversed starting from a general form of RBM energy (1).

- $P(\boldsymbol{v}) = \int \mathrm{d}\boldsymbol{h} P(\boldsymbol{v},\boldsymbol{h}) = \int \mathrm{d}\boldsymbol{h} \frac{1}{\mathcal{Z}} e^{-E(\boldsymbol{v},\boldsymbol{h})}$ and $P(\boldsymbol{v}) = \frac{1}{\mathcal{Z}} e^{-E(\boldsymbol{v})}$
- $$\begin{aligned} E(\boldsymbol{v}) &= -\log \int \mathrm{d}\boldsymbol{h} e^{-E(\boldsymbol{v},\boldsymbol{h})} \\ &= -\sum_i a_i(v_i) - \sum_\mu \log \int \mathrm{d}h_\mu e^{\frac{1}{2}\sum_\mu h_\mu^2 + \sum_{i\mu} w_{i\mu} v_i h_\mu} \end{aligned}$$
- Using the cumulant generating function and the hidden units distribution

$$\begin{aligned} K_\mu &= \log \int \mathrm{d}h_\mu q_\mu(h_\mu) e^{t h_\mu} \\ &= \sum_n k_\mu^{(n)} \frac{t^n}{n!} \end{aligned}$$

with

$$k_\mu^{(n)} = \partial_t^n K_\mu \Big|_{t=0}$$

$$q_\mu(h_\mu) = \frac{1}{\mathcal{Z}} e^{\mathcal{U}_\mu(h_\mu)}$$

- $E(\boldsymbol{v}) = -\sum_i a_i(v_i) - \sum_i \left(\sum_\mu k_\mu^{(1)} w_{i\mu}\right) v_i - \frac{1}{2}\sum_{ij}\left(\sum_\mu k_\mu^{(2)} w_{i\mu} w_{j\mu}\right) v_i v_j + \ldots$

# Limits of Hopfield model

Since Hopfield networks realize only a pair interaction between visible units it is easy to understand its limit in the learning procedure. Here is results of an argument based on a statistical mechanics approach

Hopfield Hamiltonian:
$H(\boldsymbol{v}) = -\frac{1}{2}\sum_{ij,i\neq j}\frac{1}{N}\sum_{\mu}\xi_i^{\mu}\xi_j^{\mu}v_iv_j$

$$\begin{cases} f = \frac{1}{2}\boldsymbol{m}^2 - \frac{1}{\beta}\left\langle \log[2\cosh(\beta\,\boldsymbol{m}\cdot\boldsymbol{\xi_i})]\right\rangle \\ \boldsymbol{m} = \left\langle \boldsymbol{\xi_i}\tanh(\beta\,\boldsymbol{m}\cdot\boldsymbol{\xi_i})]\right\rangle \end{cases}$$

Overlap or magnetization:
$m_\mu = \frac{1}{N}\sum_i v_i\xi_i^{\mu}$

Distribution of i.i.d. weights

$$\xi_i^{\mu} = \begin{cases} +1 & \text{with prob } \frac{1}{2} \\ -1 & \text{with prob } \frac{1}{2} \end{cases}$$

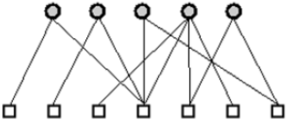
It is the case of a small number of patterns: $\alpha = \frac{M}{N} << 1$

# Hopfield phase transition

Two different solutions

Above $T \sim 1$

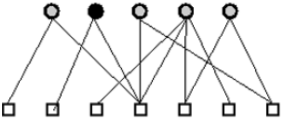
$$\begin{cases} f = -T\log(2) \\ \boldsymbol{m} = 0 \end{cases}$$



Spin-glass phase

Around $T \sim 0$

$$\begin{cases} f = -\frac{1}{2}\boldsymbol{m}^2 \\ \boldsymbol{m} = \left\langle \boldsymbol{\xi_i}\, \text{sgn}(\boldsymbol{m}\cdot\boldsymbol{\xi_i}) \right\rangle \end{cases}$$



Ferromagnetic phase

In particular $\boldsymbol{m}^2$ is bounded: $\boldsymbol{m}^2 \leq 1$ and the minimal states are of the kind $\boldsymbol{m} = (1, 0, 0, ..., 0)$

# Many patterns

Free energy $f = -\frac{1}{N\beta} < logTr(e^{-\beta H}) >$ with $\alpha$ finite leads to different saddle point equations: $f$

$$\begin{cases} \boldsymbol{m} = \langle \xi_i \tanh[\beta(\alpha r)^{1/2} z + \boldsymbol{m} \cdot \boldsymbol{\xi_i}] \rangle \\ q = \langle \tanh^2[\beta(\alpha r)^{1/2} z + \boldsymbol{m} \cdot \boldsymbol{\xi_i}] \rangle \\ r = q[1 - \beta(1-q)]^{-2} \end{cases} \xrightarrow[\text{T} \to 0]{} \begin{cases} m = \text{erf}(\frac{m}{\sqrt{2\alpha r}}) \\ q = 1 \\ r = (1-C)^{-2} \end{cases}$$

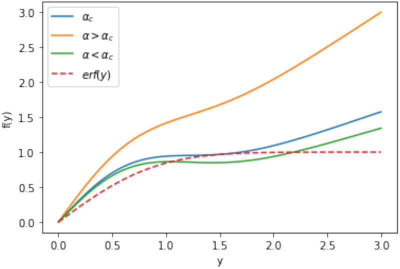
$$C = \sqrt{\frac{2}{\pi \alpha r}} e^{-\frac{m^2}{2\alpha r}}$$

Using the change of variables:

- $y = \frac{m}{\sqrt{2\alpha r}}$

the magnetization equation gives :

- $y(\sqrt{2\alpha} + \frac{2}{\sqrt{\pi}} e^{-y^2}) = \text{erf}(y)$



Critical capacity $\alpha_c \approx 0.138$

# Connection with Gaussian RBM

- The Hopfield network can be recasted in a RBM with a Gaussian hidden layer: $\mathcal{U}(h) = \frac{h^2}{2}$, with zero visible strength: $g = 0$
- The pattern retrieval can be reinterpreted in an RBM optics
- Consider Bernoulli visible units $v_i \in \{-1, 1\}$ with step activation funcion (Tab.1)
- $w_{i\mu} = \frac{1}{\sqrt{N}}\xi_{i\mu}$ and $\xi_{i\mu} = \begin{cases} +1 & \frac{p}{2} \\ -1 & \frac{p}{2} \\ 0 & 1-p \end{cases}$ where $p$ is the sparsity parameter. Here $p = 1$.
- $I_\mu(\boldsymbol{v}) = \sqrt{N} m_\mu(\boldsymbol{v})$ then the most probable value $h^*_\mu$ is $h^*_\mu = I_\mu$

# Pattern retrival

- Supose $\boldsymbol{v}$ is a (single) memory pattern: $\boldsymbol{v} = \boldsymbol{\xi_1}$
- $h_1^* = I_1 = \sqrt{N} m_1(\boldsymbol{\xi_1}) = \sqrt{N}$
- $h_\mu^* = \sqrt{N} \sum_i \frac{\xi_{i1}\xi_{i\mu}}{N}\Big|_{\mu \neq 1} \sim \mathcal{N}(0,1)$
- From Gibbs sampling:

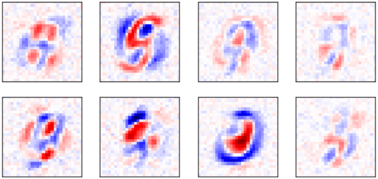
  $\boldsymbol{v} \to \boldsymbol{h} = (\sqrt{N}, h_2^*, ...)$

  $\boldsymbol{h} \to v_i = \Theta\big(\xi_{i1} + \sqrt{N} \sum_{\mu \neq 1}^{M} \xi_{i\mu} h_\mu^*\big)$
- In order to retrieve the initial pattern the second contribution needs to be less than 1. This is obtained by $\frac{M}{N} < 1$.
  In the limit $N \to \infty$ we need e.g. M to be finite

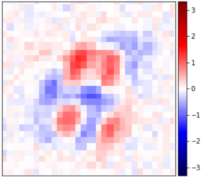
Choosing a non linear transfer function such as ReLU is a simple way to suppress undesired inputs.

# Simulations on MNIST

A simulation of an RBM-Gaussian machine with $\alpha \approx 0.26$ shows that weights have a "chaotic" behaviour.

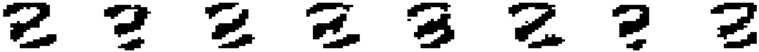


RBM-Gaussian weights



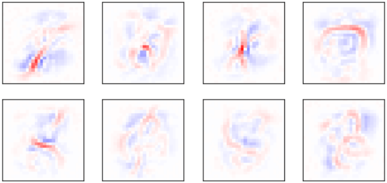
Many interactions with similar strenght

Many weights compete among each other and the resulting generative power of the machine decrease

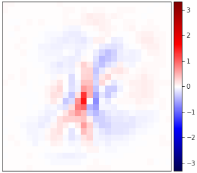


# ReLU-RBM simulation on MNIST

The result of a ReLU-RBM with the same number of hidden units substantially different

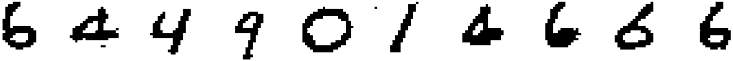


RBM-ReLU weights



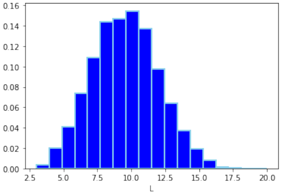
Sparse connection with few stronger

Different combination of the (stronger) learned features are combined to produce different variants of the same digits. Manyof those are not contained in the training set

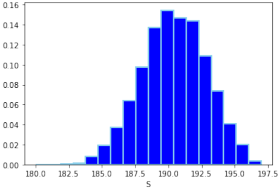


# Strongly activated hidden units

In each generated handwritten digit image lots of hidden units are silent, while the remaining has largely varying activations, some weak and few very strong.



Strongly activated hidden units ($L$)



Silent hidden units ($S$)

$L$ can be estimated by the participation ratio:

$$\hat{L} = [(\sum_\mu h_\mu)^a)^2/(\sum_\mu h_\mu^{2a})] = PR_a(h)$$

The value of $a$ is set to 3.

Starting from a typical configuration $\boldsymbol{h}$:

$$h_\mu = \begin{cases} m\sqrt{N} & \text{if} \quad 1 \leq \mu \leq L \\ \sqrt{r} x_\mu & \text{if} \quad L+1 \leq \mu \leq M \end{cases}$$

$$PR_3(h) \sim \frac{(Lm^3 N^{3/2} + (N-L) r^{3/2})^2}{Lm^6 N^3 + (N-L) r^3}$$

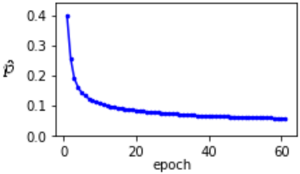
$$= L \times \frac{(1 + \frac{(N-L)}{N^{3/2}} \frac{r^{3/2}}{Lm^3})^2}{1 + \frac{(N-L)}{N^3} \frac{r^3}{Lm^6}} \xrightarrow{N \to \infty} L$$

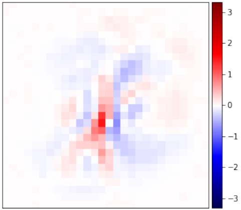
**Compositional phase**

The machine finish the training with more than one strongly activated hidden unit

# Weight sparsity

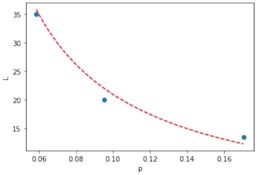
- After training the machine reach a low degree of sparsity



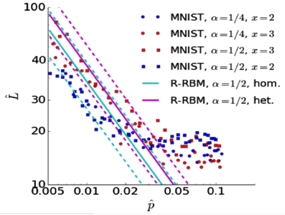


Weight sparsity: $\hat{p} = \frac{1}{MN}\sum_{\mu}[(\sum_i w_{i\mu}^2)/(\sum_i w_{i\mu}^4)]$

- Different RBMs has reach different sparsity values. This affects also $L$



$$L \sim \ell^*/p$$



J.Tubiana, R.Monasson (2017)

# Replica theory

Studies related to Random RBMs (R-RBMs) support the observations.

- Assumption:
  - system state $(E1, E2, ..., E_N)$
  - system distribution $\mu_B(j) = \frac{1}{Z} e^{-\beta E_j}$
  - $E_i \quad i.i.d. \sim \; e.g. \, \mathcal{N}(0, \sigma^2)$
- The Replica trick allows to obtain easily $\mathbb{E} \, logZ \sim f$
- Instead of computing $\mathbb{E} \, logZ$ it results easyer computing $\mathbb{E} \, Z^n$
- $Z^n = \text{exp}[n \log Z] \overset{n<<1}{\simeq} 1 + n \log Z \Rightarrow \log Z \overset{n<<1}{\simeq} \frac{Z^n - 1}{n}$
- $\mathbb{E} \log Z = \lim_{n \to 0} \frac{\mathbb{E}(Z^n) - 1}{n}$

# Replica theory

- R-RBM ensamble are characterized by random patterns

$$w_{i\mu} = \begin{cases} +1/\sqrt{N} & \frac{p_i}{2} \\ -1/\sqrt{N} & \frac{p_i}{2} \\ 0 \quad 1 - p_i \end{cases} \qquad p_i \in [0, 1]$$

- Energy of the general bipartite network

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\sum_{i=1}^{N} \sum_{\mu=1}^{M} w_{i\mu} v_i h_\mu +$$

$$-\sum_{i=1}^{N} \mathcal{U}_v(v_i) + \sum_{\mu=1}^{M} \mathcal{U}_h(h_\mu)$$

- Partition function

$$Z = = \sum_{\{\boldsymbol{v}\}} \sum_{\{\boldsymbol{h}\}} e^{-\beta E(\boldsymbol{v}, \boldsymbol{h})}$$

- Replica partion function

$$Z^n = \int (\prod_{i,a} \mathrm{d}v_{i,a}) \times$$

$$\times \int (\prod_{\mu,a} \mathrm{d}h_{\mu,a}) e^{-\beta E(\boldsymbol{v}_a, \boldsymbol{h}_a)}$$

# Towards free energy

- Simulations allow to make an ansatz for the hidden units: fixing e.g. the first $L$ to be the strongly activated ones $h_1^a = h_2^a = ... = h_L^a = m\sqrt{N}$

$$Z^n = \int (\prod_{i,a} \mathrm{d}v_{i,a}) \int (\prod_{\mu,a} \mathrm{d}h_{\mu,a}) \,\mathsf{exp}\Big[ - \sum_a \beta\, L\, \mathcal{U}_h(m\sqrt{N}) - \sum_{a,\mu>L} \beta\, \mathcal{U}_h(h_\mu^a)$$

$$-\beta\, m\sqrt{N}\, L \sum_{i,\mu=1,..L} w_{i\mu} \sum_a v_i^a - \beta \sum_{i,\mu>L} w_{i\mu} \sum_a v_i^a h_\mu^a \Big]$$

- The average of $Z^n$ on the last term brings to a quartic interaction term $\sim v_i^a\, v_i^b\, h_\mu^a\, h_\mu^b$.The decoupling is done by the HS transformation that produce new fields

$$\bar{Z}^n = \prod_{a\leq b} \frac{\mathrm{d}\bar{q}^{ab}\mathrm{d}q^{ab}}{2N\,\beta} \,\mathsf{exp}\Big[ - \beta\, N \sum_{a\leq b} \bar{q}^{ab} q^{ab} - \beta\, n\, L\, \mathcal{U}_h(m\sqrt{N}) \Big] \times$$

$$\times \prod_i \prod_a \int \mathrm{d}v_i^a \,\mathsf{exp}\Big[ - \beta \sum_a \mathcal{U}_v(v_i^a) + \beta \sum_{a\leq b} \bar{q}^{ab} v_i^a v_i^b \tfrac{p_i}{p} + \beta\, m \big( \sum_\mu^L \sqrt{N} w_{i\mu} \big) \big( \sum_a v_i^a \big) \Big] \times$$

$$\times \Big( \prod_a \int \mathrm{d}h^a \,\mathsf{exp}\Big[ - \beta \sum_a \mathcal{U}_h(h^a) + \tfrac{\beta^2 p}{2} \sum_{a\leq b} q^{ab} h^a h^b \tfrac{p_i}{p} \Big] \Big)^{\alpha\, N - L}$$

# Towards free energy

- Functional form of $\bar{Z}^n = \int \mathrm{d}a\, \mathrm{d}b, ... e^{-\beta F(a,b,...)}$

- Saddle-point approximation $\bar{Z}^n \approx$ exp$(-\beta F(a,b,...)\big|_{a^*,b^*,...}$
  $(a^*, b^*, ...)$ are the saddle point value of $F$

- $\bar{Z}^n \approx 1 - \beta F(a,b,...)\big|_{a^*,b^*,...}$

- Free energy:
  $$f = \frac{L\, m^2}{2} + \frac{\alpha}{2}(r\, C + B\, q) + \alpha \frac{1}{N} \sum_i \langle \int Dz \min_v \Big[ \mathcal{U}_v(v) - (m\, W + \\ + z\sqrt{\alpha \pi r})v - \frac{\alpha}{2} B \frac{p_i}{p} v^2 \Big] \rangle_W + \alpha \int Dz \min_h \Big[ \mathcal{U}_h(h) - \frac{C}{2} h^2 - z\sqrt{\alpha p\, r} \Big]$$

where

$$\begin{cases} W = \sqrt{N} \sum_\mu w_{i\mu} \\ p = \sum_i \frac{p_i}{N} \end{cases}$$

Replica symmetric ansatz

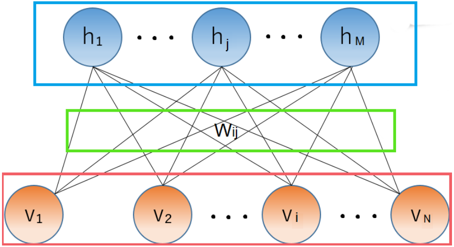
$$\begin{cases} q^{ab} = q + \delta_{ab} \frac{C}{p\beta} \\ \bar{q}^{ab} = \frac{\alpha \beta p}{2} \left[ 2r(1 - \delta_{ab}) + \delta_{ab}(r + \frac{B}{p\beta}) \right] \end{cases}$$

# Free energy and RBM structure

$$f = \frac{L\,m^2}{2} + \alpha \int Dz \min_h \Big[\mathcal{U}_h(h) - \frac{C}{2}h^2 -$$

$$z\sqrt{\alpha p\, r}\Big]$$

$$+\frac{\alpha}{2}(r\,C + B\,q)$$

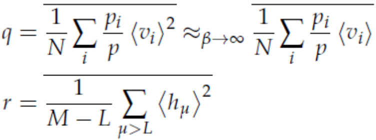


$$+\alpha\frac{1}{N}\sum_i \langle \int Dz \min_v \Big[\mathcal{U}_v(v) - (m\,W$$

$$+z\sqrt{\alpha\pi r})v - \frac{\alpha}{2}B\frac{p_i}{p}v^2\Big]\rangle_W$$

# Parameter interpretation

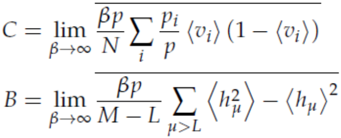
The saddle point equations (e.g. $\frac{\partial(-\beta F/N)}{\partial q^{ab}} = 0$) give insights about the interpretation of the saddle point parameters



$$q = \overline{\frac{1}{N}\sum_i \frac{p_i}{p}\langle v_i\rangle^2} \approx_{\beta\to\infty} \overline{\frac{1}{N}\sum_i \frac{p_i}{p}\langle v_i\rangle}$$

$$r = \overline{\frac{1}{M-L}\sum_{\mu>L}\langle h_\mu\rangle^2}$$

J.Tubiana, R.Monasson (2017)



$$C = \lim_{\beta\to\infty} \overline{\frac{\beta p}{N}\sum_i \frac{p_i}{p}\langle v_i\rangle(1-\langle v_i\rangle)}$$

$$B = \lim_{\beta\to\infty} \overline{\frac{\beta p}{M-L}\sum_{\mu>L}\left\langle h_\mu^2\right\rangle - \langle h_\mu\rangle^2}$$

J.Tubiana, R.Monasson (2017)

- $(q, r, C, B)$ are respectivelly the (weighted) mean activity in the visible layer, the square average activation of the weakly activated hidden units, the rescaled variance of the visible and hidden units.

# Homogeneus free energy

- Solution for the homogeneus case ($p_i \equiv p$), Bernoulli visible units with constant strenght ($g_i \equiv g$) and ReLU hidden units:
- $f_{GS}(L,m,r,B,q,C) = \frac{L\,m^2}{2} + \frac{\alpha}{2}(r\,C + B\,q) - \sqrt{\alpha p\,r}\langle H^{(1)}\Big( - \Big[\frac{g+m\,W+\alpha B/2}{\sqrt{\alpha p\,r}}\ \Big]\Big)\rangle_W + -\frac{\alpha pq}{2(1-C)}\,H^{(2)}\Big(\frac{\theta}{\sqrt{pq}}\Big)$

  where $H^{(k)}(x) = \int_x^{\infty} Dz(z-x)^k$ and $Dz$ is the Gaussian measure
- Saddle point equations:

$$m = \frac{1}{L}\langle W\,H^{(0)}\Big( - \Big[\frac{g+mW+\alpha B/2}{\sqrt{\alpha pr}}\Big]\ \Big)\rangle_W \qquad r = \frac{pq}{(1-C)^2}H^{(2)}\Big(\frac{\theta}{\sqrt{pq}}\Big)$$

$$B = \frac{p}{(1-C)}H^{(0)}\Big(\frac{\theta}{\sqrt{pq}}\Big)$$

$$q = \langle H^{(0)}\Big( - \Big[\frac{g+mW+\alpha B/2}{\sqrt{\alpha pr}}\Big]\ \Big)\rangle_W$$

$$C = \frac{\sqrt{p}}{\sqrt{2\pi\alpha r}}\langle e^{-\frac{1}{2}\frac{-(g+mW+\alpha B/2)^2}{\alpha pr}}\rangle_W$$

# Homogeneus free energy

- Using Feynman integration techniques the energy and the equations can be also recast in a different form:

$$f_{GS}(L,m,r,B,q,C) = \frac{L\,m^2}{2} + \frac{\alpha}{2}(r\,C + B\,q) - \sqrt{\alpha p\,r}\Big\langle \frac{1}{2} e^{-\frac{1}{2}(\frac{(g+m\,W+\alpha B/2)^2}{\alpha\,p\,r}} +$$

$$+\frac{1}{2}\Big(\frac{g+m\,W+\alpha B/2}{\sqrt{\alpha p\,r}}\Big)\Big(1+\mathsf{erf}\big(\frac{g+m\,W+\alpha B/2}{\sqrt{2\alpha p\,r}}\big)\Big)\Big\rangle_W - \frac{\alpha pq}{2(1-C)}\,\frac{1}{2}e^{-\frac{1}{2}\frac{\theta^2}{pq}}\Big[-\sqrt{\frac{2}{\pi}}\frac{\theta}{\sqrt{pq}} +$$

$$+e^{\frac{1}{2}\frac{\theta^2}{pq}}(1+\frac{\theta^2}{pq})\Big(1+\mathsf{erf}\big(\frac{\theta}{\sqrt{2pq}}\big)\Big)\Big]$$

$$m = \frac{1}{L}\Big\langle W\,\frac{1}{2}\Big(1+\mathsf{erf}\big(\frac{g+m\,W+\alpha B/2}{\sqrt{2\alpha p\,r}}\big)\Big)\Big\rangle_W$$

$$q = \Big\langle \frac{1}{2}\Big(1+\mathsf{erf}\big(\frac{g+m\,W+\alpha B/2}{\sqrt{2\alpha p\,r}}\big)\Big)\Big\rangle_W$$

$$C = \frac{\sqrt{p}}{\sqrt{2\pi\alpha r}}\langle e^{-\frac{1}{2}\frac{-(g+mW+\alpha B/2)^2}{\alpha p r}}\rangle_W$$

$$r = \frac{pq}{(1-C)^2}\,\frac{1}{2}e^{-\frac{1}{2}\frac{\theta^2}{pq}}\Big[-\sqrt{\frac{2}{\pi}}\frac{\theta}{\sqrt{pq}} +$$

$$+e^{\frac{1}{2}\frac{\theta^2}{pq}}(1+\frac{\theta^2}{pq})\Big(1+\mathsf{erf}\big(\frac{\theta}{\sqrt{2pq}}\big)\Big)\Big]$$

$$B = \frac{p}{(1-C)}\,\frac{1}{2}\,\mathsf{erfc}\big(\frac{\theta}{\sqrt{2pq}}\big)$$

# Hopfield network with threshold

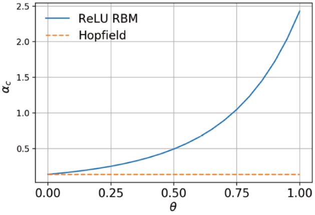
- The Hopfield model can be reproduced but with the presence of a threshold $(g = -\alpha B/2, L = 1, p = 1)$

$$\begin{cases} m = \frac{1}{2}\text{erf}\left(\frac{m}{\sqrt{2\alpha r}}\right) \\ q = 1/2 \\ C = \frac{1}{\sqrt{2\pi\alpha r}} e^{-\frac{m^2}{2\alpha r}} \\ r = \frac{1}{2(1-C)^2} H^{(2)}(\sqrt{2}\theta) \\ B \quad \text{decoupled from the previous} \end{cases}$$

$$\longrightarrow$$

$$\begin{array}{c} (m, r, C, \alpha) \\ \downarrow \\ (\tilde{m}/2, \tilde{r}H^{(2)}/2, \tilde{C}, \tilde{\alpha}/(2H^{(2)})) \end{array}$$

$$\begin{cases} m = \frac{1}{2}\text{erf}\left(\frac{\tilde{m}}{\sqrt{2\tilde{\alpha}\tilde{r}}}\right) \\ \tilde{C} = \sqrt{\frac{2}{\pi\tilde{\alpha}\tilde{r}}} e^{-\frac{m^2}{2\tilde{\alpha}\tilde{r}}} \\ \tilde{r} = (1 - \tilde{C})^{-2} \end{cases}$$

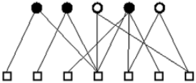


$$\alpha_c = \frac{\tilde{\alpha}_{c,Hpf}}{2H^{(2)}(\sqrt{2}\theta)}$$

- The threshold $(\theta)$ can be increased to escape from the glassy phase of the Hopfield model.
  The machine can learn again with an higher capacity $\alpha_c$

# Compositional phase

- Case of simulations: $L \sim \frac{l}{p}$ for $p << 1$.
  More than one hidden unit cooperate to generate data.



Compositional phase

- Change of parameters:

$$\begin{cases} L = l/p \\ \theta = \tilde{\theta}\sqrt{p} \\ m = \tilde{m}\frac{p}{2} \\ g = \tilde{g}p \end{cases} \qquad \begin{cases} r = \tilde{r}p \\ B = \tilde{B}p \\ f = \tilde{f}p \end{cases}$$

$$\begin{cases} M = \frac{1}{2}\frac{\tilde{m}}{\sqrt{\alpha\tilde{r}}} \\ \theta_v = -\frac{\tilde{g}+\alpha\tilde{B}/2}{\sqrt{\alpha\tilde{r}}} \\ \theta_h = \tilde{\theta}/\sqrt{q} \end{cases}$$

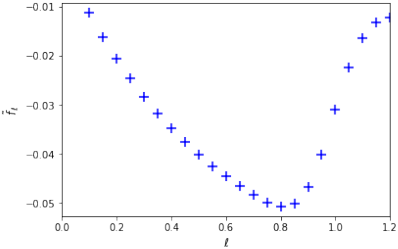
- Saddle point equations:

$$\begin{cases} \tilde{m} = \frac{2}{l}\langle W H^{(0)}\Big( -(MW - \theta_v) \Big)\rangle_W \\ C = \frac{1}{\sqrt{2\pi\alpha\tilde{r}}}\langle e^{-\frac{1}{2}(MW-\theta_v)^2}\rangle_W \\ q = \langle H^{(0)}\Big( -(MW - \theta_v) \Big)\rangle_W \\ \tilde{r} = \frac{q}{(1-C)^2}H^{(2)}(\theta_h) \\ \tilde{B} = \frac{1}{(1-C)}H^{(0)}(\theta_h) \end{cases}$$

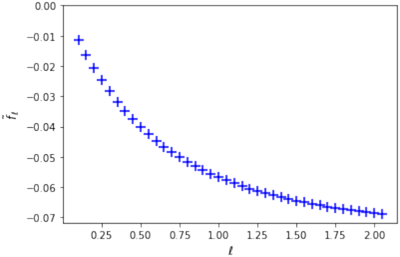
# Rescaled free energy

The new saddle point equations and the parameters give the rescaled free energy:

- $\tilde{f}_{GS} = -\frac{l}{8}\tilde{m}^2 - \frac{\alpha}{2}(\tilde{r}C + \tilde{B}q) - \tilde{g}q + \frac{\alpha\sqrt{q}}{2}\,\tilde{\theta}\,\frac{H^{(1)}\left(\frac{\tilde{\theta}}{\sqrt{q}}\right)}{1-C}$



$\tilde{g} = -0.21,\ \alpha = 0.5,\ \tilde{\theta} = 1.5$



$\tilde{g} = -0.17,\ \alpha = 0.5,\ \tilde{\theta} = 1.5$

# Bibliography

[1] J. Tubiana and R. Monasson, "Emergence of compositional representations in restricted boltzmann machines," *Physical review letters*, vol. 118, no. 13, p. 138 301, 2017.

[2] D. J. Amit, H. Gutfreund, and H. Sompolinsky, "Storing infinite numbers of patterns in a spin-glass model of neural networks," *Physical Review Letters*, vol. 55, no. 14, p. 1530, 1985.

[3] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, "A high-bias, low-variance introduction to machine learning for physicists," *Physics Reports*, vol. 810, pp. 1–124, May 2019, ISSN: 0370-1573. DOI: 10.1016/j.physrep.2019.03.001. [Online]. Available: http://dx.doi.org/10.1016/j.physrep.2019.03.001.

[4] T. Castellani and A. Cavagna, "Spin-glass theory for pedestrians," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 05, P05012, 2005.

[5] D. J. Amit, H. Gutfreund, and H. Sompolinsky, "Spin-glass models of neural networks," *Physical Review A*, vol. 32, no. 2, p. 1007, 1985.

# Bibliography (cont.)

[6] P. Dayan and L. F. Abbott, *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Computational Neuroscience Series, 2001.

[7] J. Son, “Replica trick on spin glasses and boolean satisfiability,”, 2018.

[8] J. Tubiana, “Restricted boltzmann machines: From compositional representations to protein sequence analysis,” Ph.D. dissertation, PSL Research University, 2018.

[9] D. J. MacKay and D. J. Mac Kay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[10] M. Mezard and A. Montanari, *Information, physics, and computation*. Oxford University Press, 2009.

[11] J. Schlüter, “Restricted boltzmann machine derivations,” Technical Report TR-2014-13, Österreichisches Forschungsinstitut für . . ., Tech. Rep., 2014.

# Some questions on RBM

Difference in the conditional probability between $\{0,1\}$ Bernoulli RBMs and $\{-1,1\}$ Bernoulli RBMs

- $\{0,1\}$ Bernoulli RBMs
- $P(h_{\tilde{\mu}}=1|\boldsymbol{v}) = P(\boldsymbol{v}, h_{\tilde{\mu}}=1)/P(\boldsymbol{v})$

$$\frac{\sum\limits_{\{h_\mu, h_{\tilde{\mu}}=1\}} P(\boldsymbol{v},\boldsymbol{h})}{\sum_{\{h_\mu\}} P(\boldsymbol{v},\boldsymbol{h})} = \frac{\sum\limits_{\{h_\mu, h_{\tilde{\mu}}=1\}} e^{-E(\boldsymbol{v},\boldsymbol{h})}}{\sum_{\{h_\mu\}} e^{-E(\boldsymbol{v},\boldsymbol{h})}}$$

$$= \sum_{\{h_\mu, h_{\tilde{\mu}}=1\}} e^{-E(\boldsymbol{v},\boldsymbol{h})} \Bigg/ \Bigg( \sum_{\{h_\mu, h_{\tilde{\mu}}=1\}} e^{-E(\boldsymbol{v},\boldsymbol{h})} + \sum_{\{h_\mu, h_{\tilde{\mu}}=0\}} e^{-E(\boldsymbol{v},\boldsymbol{h})} \Bigg)$$

$$= \frac{1}{1 + \frac{\sum\limits_{\{h_\mu, h_{\tilde{\mu}}=0\}} e^{-E(\boldsymbol{v},\boldsymbol{h})}}{\sum\limits_{\{h_\mu, h_{\tilde{\mu}}=1\}} e^{-E(\boldsymbol{v},\boldsymbol{h})}}} = \frac{1}{1 + \frac{\sum\limits_{\{h_\mu, h_{\tilde{\mu}}=0\}} e^{-E(\boldsymbol{v},\boldsymbol{h})}}{\sum\limits_{\{h_\mu, h_{\tilde{\mu}}=1\}} e^{-E(\boldsymbol{v},\boldsymbol{h})}}}$$

# Some questions on RBM

$$= \left( 1 + \frac{\sum\limits_{\{h_\mu, h_{\tilde{\mu}}=0\}} \exp(\sum_{i,\mu} v_i\, w_{i\mu}\, h_\mu + \sum_i a_i v_i + \sum_\mu b_\mu\, h_\mu)}{\sum\limits_{\{h_\mu, h_{\tilde{\mu}}=1\}} \exp(\sum_{i,\mu} v_i\, w_{i\mu}\, h_\mu + \sum_i a_i v_i + \sum_\mu b_\mu\, h_\mu)} \right)^{-1}$$

- Separating, inside the energy, the terms independent from $h_{\tilde{\mu}}$:

$$= \left( 1 + \frac{\exp(\sum_i v_i\, w_{i\tilde{\mu}}\, h_{\tilde{\mu}} + b_{\tilde{\mu}} h_{\tilde{\mu}})\big|_{h_{\tilde{\mu}}=0}}{\exp(\sum_i v_i\, w_{i\tilde{\mu}}\, h_{\tilde{\mu}} + b_{\tilde{\mu}} h_{\tilde{\mu}})\big|_{h_{\tilde{\mu}}=1}} \right)^{-1} = \Big(1 + \exp[-(b_{\tilde{\mu}} + \sum_i v_i w_{i\tilde{\mu}})\,]\Big)^{-1}$$

$$= \sigma(b_{\tilde{\mu}} + \sum_i v_i w_{i\tilde{\mu}})$$

- Doing the same steps for the $\{-1, 1\}$ Bernoulli RBM:

$$= \left( 1 + \frac{\exp(\sum_i v_i\, w_{i\tilde{\mu}}\, h_{\tilde{\mu}} + b_{\tilde{\mu}} h_{\tilde{\mu}})\big|_{h_{\tilde{\mu}}=-1}}{\exp(\sum_i v_i\, w_{i\tilde{\mu}}\, h_{\tilde{\mu}} + b_{\tilde{\mu}} h_{\tilde{\mu}})\big|_{h_{\tilde{\mu}}=1}} \right)^{-1} = \Big(1 + \exp[-2(b_{\tilde{\mu}} + \sum_i v_i w_{i\tilde{\mu}})\,]\Big)^{-1}$$

$$= \sigma(2[b_{\tilde{\mu}} + \sum_i v_i w_{i\tilde{\mu}}])$$

# Learning procedure

- The log-likelihood is $\mathcal{L} = \text{log} P(\boldsymbol{v}; \boldsymbol{\theta})$

$$\frac{\partial \text{log} P(\boldsymbol{v},\boldsymbol{\theta})}{\partial \theta} = \sum_{\{\boldsymbol{h}\}} P(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta}) \frac{\partial(-E(\boldsymbol{v},\boldsymbol{h},\boldsymbol{\theta}))}{\partial \theta} - \sum_{\{\boldsymbol{v}\}} P(\boldsymbol{v}; h) \sum_{\{\boldsymbol{h}\}} P(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta}) \frac{\partial(-E(\boldsymbol{v},\boldsymbol{h},\boldsymbol{\theta}))}{\partial \theta}$$

- Defining $G_\theta(\boldsymbol{x},\boldsymbol{\theta}) = \sum_{\{\boldsymbol{h}\}} P(\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta}) \frac{\partial(-E(\boldsymbol{x},\boldsymbol{h},\boldsymbol{\theta}))}{\partial \theta}$

  (Note: we are deriving over one parameter $\theta$ of all the possible set $\boldsymbol{\theta}$)

$$\frac{\partial \text{log} P(\boldsymbol{v},\boldsymbol{\theta})}{\partial \theta} = G_\theta(\boldsymbol{v},\boldsymbol{\theta}) - \big\langle G_\theta(\boldsymbol{x},\boldsymbol{\theta}) \big\rangle_{P(\boldsymbol{v};\boldsymbol{\theta})}$$

- The second term is averaged over the model distribution ($P(\boldsymbol{v};\boldsymbol{\theta})$)

# Weights Learning procedure

- $$G_{w_{i\mu}}(\boldsymbol{v},\boldsymbol{\theta}) = \sum_{\{\boldsymbol{h}\}} P(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta}) \frac{\partial(-E(\boldsymbol{v},\boldsymbol{h},\boldsymbol{\theta}))}{\partial w_{i\mu}}$$
  $$= \sum_{\{\boldsymbol{h}\}} P(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta})\; v_i h_\mu$$
  $$= v_i \sum_{\{\boldsymbol{h}\}} P(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta})\; h_\mu$$
- Using ”Conditional independence” property:
  $$= v_i \sum_{\{h_\mu\}} \sum_{\{\boldsymbol{h}\neq h_\mu\}} P(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta}) P(h_\mu|\boldsymbol{v};\boldsymbol{\theta})\; h_\mu$$
  $$= v_i \sum_{\{h_\mu\}} P(h_\mu|\boldsymbol{v};\boldsymbol{\theta})\; h_\mu$$
- In the $\{0,1\}$ Bernoulli RBM case
  $$= v_i \Big( P(h_\mu = 1|\boldsymbol{v};\boldsymbol{\theta}) \cdot 1 + P(h_\mu = 0|\boldsymbol{v};\boldsymbol{\theta}) \cdot 0 \Big)$$
  $$= \boxed{v_i \sigma(b_{\tilde{\mu}} + \sum_i v_i w_{i\tilde{\mu}}) \overset{?}{=} v_i h_\mu} \rightarrow \sigma \in [0,1] \neq \{0,1\}$$

# Probability distribution issue

- Probability distribution of the weights for the R-RBMs case

$$\mathsf{w}_{i\mu} = \begin{cases} +1/\sqrt{N} & \frac{p_i}{2} \\ -1/\sqrt{N} & \frac{p_i}{2} \\ 0 \quad 1-p_i \end{cases} \qquad p_i \in [0,1]$$

- After the rescaled parametrization, where in particular $L = \ell/p$, the variable $W$ became $W_i = \sum_{\mu}^{\ell/p} w_{i\mu}$
- $W$ is a sum over $\ell/p$ terms with a fraction $p_i = x_i p$ (homogeneus case $p_i = p$) non zero.
- In the $p \to 0$ limit the distribution of the random variable $W$ is:

$$P_{\ell x_i}(W_i = w \in \mathbb{Z}) = e^{-lx_i} I_w(lx_i)$$

$$I_\alpha(x) = \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(k+\alpha+1)} (\frac{x}{2})^{2k+\alpha}$$