Artificial Intelligence and machine learning. Challenges and opportunities for mathematicians

Pierluigi Contucci

University of Bologna

pierluigi.contucci@unibo.it

A perspective on Artificial Intelligence in Industry and Research, CAE Conference

October 29th, 2019

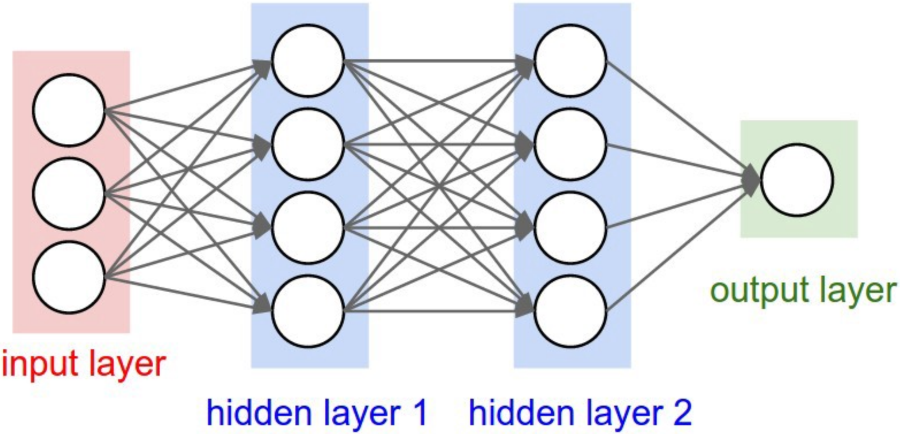
Overview

1. Introduction
   - The AI realm, machine learning and deep learning
   - Why mathematical (and physical!) approaches
   - An inverse problem in a complex landscape
   - Statistical Physics perspective
2. The model: Deep Boltzmann Machines
   - Definitions
   - Mathematical quantities
   - Theorems: annealing and replica symmetry
   - Architectural constraints
3. What's next?
   - ...
   - ...

Introduction

- Machine learning, classical AI (Symbolic)
- High dimensional problems (low dimensional)
- Deep architectures in machine learning

Introduction



Introduction

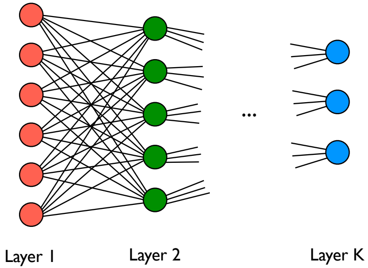
- Deep supervised learning: statistical mechanics inverse problem with assigned boundary conditions
- Euristically: non convex problem, very large (high entropy!) local minima
- Use a class of tractable models (Boltzmann Machines) and aim at their exact solutions
- Investigate their architectural constraints
- ...

Definition of the DBM

We consider a statistical mechanics model composed by

- N binary Ising spins
- arranged over K layers $L_1, \ldots, L_K$ of sizes $N_1, \ldots, N_K$ respectively with $\sum_{p=1}^{K} N_p = N$
- spins in the layer L_p interact with all those in the layer L_{p-1}, L_{p+1} and only with them
- the weights connecting layers L_p and L_{p+1} are $N_p \times N_{p+1}$ real valued i.i.d. random couplings sampled from a standard Gaussian distribution

Definition of the DBM



Schematic representation of a DBM with K layers. Each circle represents a spin variable while all the interactions are drawn among spins in adjacent layers (but there are no intra-layer interactions)

Themodynamic limit and form factors

We will focus on the properties of the DBM in the thermodynamic limit, namely when $N \to \infty$ so

- we denote by $\Lambda_N \equiv (N_1, \ldots, N_K)$ and we assume for every $p = 1, \ldots, K$ that the relative sizes $\frac{N_p}{N}$, that we refer to as *form factors*, of the layers converge in the large volume limit:

$$\lambda_p^{(N)} \equiv \frac{N_p}{N} \xrightarrow[N \to \infty]{} \lambda_p \in [0, 1]$$

- we denote by $\lambda = (\lambda_1, \ldots, \lambda_K)$ the relative sizes in the large volume limit. Notice that $\sum_{p=1}^{K} \lambda_p = 1$

Hamiltonian of the DBM

Definition

The random Hamiltonian (or *cost function* to keep a machine learning jargon), of a DBM is

$$H_{\Lambda_N}(\sigma) = -\frac{\sqrt{2}}{\sqrt{N}} \sum_{p=1}^{K-1} \sum_{(i,j)\in L_p \times L_{p+1}} J_{ij}^{(p)} \, \sigma_i \sigma_j$$

where $J_{ij}^{(p)}$, $(i,j) \in L_p \times L_{p+1}$, $p = 1, \ldots, K-1$ is a family of i.i.d. standard Gaussian random variables

Remark: one can also consider (random) external fields

Overlap and covariance

Notice that H_{Λ_N} is a gaussian process on $\{\pm 1\}^N$ with covariance

$$\mathbb{E}\, H_{\Lambda_N}(\sigma)\, H_{\Lambda_N}(\tau) \;=\; 2\, N \sum_{p=1}^{K-1} \lambda_p^{(N)}\, \lambda_{p+1}^{(N)}\;\; q_{L_p}(\sigma,\tau)\; q_{L_{p+1}}(\sigma,\tau)$$

where

Definition

Given two spin configurations $\sigma, \tau \in \{\pm 1\}^N$, for every $p = 1, \ldots, K$ we define the *overlap* over the layer L_p as

$$q_{L_p}(\sigma,\tau) \;=\; \frac{1}{N_p} \sum_{i \in L_p} \sigma_i\, \tau_i \;\in [-1, 1]\,.$$

Partition function and thermodynamic pressure of a DBM

Definition

Given $\beta > 0$, the random partition function is

$$Z_{\Lambda_N}(\beta) = \sum_{\sigma \in \{-1,1\}^N} e^{-\beta\, H_{\Lambda_N}(\sigma)} \,.$$

We call random pressure density the quantity $\dfrac{1}{N} \log Z_{\Lambda_N}(\beta)$

Self averaging of the pressure

Main question: properties of $\frac{1}{N}\ \log Z_{\Lambda_N}(\beta)$ as $N \to \infty$

First key property: self averaging

$$\lim_{N\to\infty} \frac{1}{N}\ \log Z_{\Lambda_N}(\beta)\ =\ \lim_{N\to\infty} p_{\Lambda_N}^{DBM}(\beta)\ \ a.s.$$

where

$$p_{\Lambda_N}^{DBM}(\beta) \equiv \frac{1}{N}\mathbb{E}\ \log Z_{\Lambda_N}(\beta)$$

is called *quenched pressure*.

Main idea

The main idea is to construct an interpolation between a DBM with K layers and K independent Sherringhton-Kirkpatrick models

Given $a = (a_p)_{1 \le p \le K-1} \in (0, \infty)^{K-1}$, for every $p = 1, \dots, K$ we consider an SK model of size N_p at inverse temperature $\beta \sqrt{\lambda_p^{(N)} \theta_p(a)}$, where we set

$$\begin{cases} \theta_1(a) \equiv a_1 \\ \theta_p(a) \equiv \dfrac{1}{a_{p-1}} + a_p \quad \text{if } p = 2, \dots, K-1 \\ \theta_K(a) \equiv \dfrac{1}{a_{K-1}} \end{cases} .$$

Main result

Theorem

The quenched pressure of the DBM satisfies the following lower bound

$$p_{\Lambda_N}^{DBM}(\beta) \geq \sum_{p=1}^{K} \lambda_p^{(N)} \, p_{N_p}^{SK}\left(\beta\sqrt{\lambda_p^{(N)}\,\theta_p(a)}\right) - \frac{\beta^2}{2}\sum_{p=1}^{K}\left(\lambda_p^{(N)}\right)^2 \theta_p(a) + \\ + \beta^2 \sum_{p=1}^{K-1} \lambda_p^{(N)} \, \lambda_{p+1}^{(N)}$$

for any choice of $a = (a_p)_{1 \leq p \leq K-1} \in (0, \infty)^{K-1}$

Main result

Corollary

$$\liminf_{N\to\infty} p_{\Lambda_N}^{DBM}(\beta) \;\geq$$

$$\sup_{a\in(0,\infty)^{K-1}} \left\{ \sum_{p=1}^{K} \lambda_p \; p^{SK}\left(\beta\sqrt{\lambda_p\,\theta_p(a)}\,\right) \;-\; \frac{\beta^2}{2}\sum_{p=1}^{K} \lambda_p^2\,\theta_p(a) \right\} +$$

$$+ \;\beta^2 \sum_{p=1}^{K-1} \lambda_p \lambda_{p+1} \;.$$

The annealed regime

Under what conditions the model is in the annealed state? Annealed means structurally convex

Consider a *DBM* with $K = 2, 3, 4$ layers and define

$$A_K = \left\{(\beta, \lambda) : 4\beta^4 \leq \phi_K(\lambda)\right\} ,$$

where we set

$$\phi_2(\lambda) \equiv \frac{1}{\lambda_1 \lambda_2}$$

$$\phi_3(\lambda) \equiv \frac{1}{\lambda_1 \lambda_2 + \lambda_2 \lambda_3}$$

$$\phi_4(\lambda) \equiv \min\{t > 0 : 1 - t\,(\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_4) + t^2\,\lambda_1\lambda_2\lambda_3\lambda_4 = 0\} .$$

The annealed regime

Theorem

If $(\beta, \lambda) \in A_K$ *then there exists*

$$\lim_{N\to\infty} p_{\Lambda_N}^{DBM}(\beta) = \lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}\, Z_{\Lambda_N}(\beta) = \log 2 + \beta^2 \sum_{p=1}^{K-1} \lambda_p \lambda_{p+1} \,.$$

The annealed regime

Some observation on the previous condition. What can we do to squeeze the annealed reagime as much as possible?

- $\beta \leq 1$ the DBM is in the annealed regime for any choice of λ.

- the infimum of $\phi_K(\lambda)$ is reached for

$$\begin{cases} \lambda_1 = \lambda_2 = \frac{1}{2} & \text{if } K = 2 \\ \lambda_2 = \frac{1}{2},\, \lambda_1 + \lambda_3 = \frac{1}{2} & \text{if } K = 3 \\ (\lambda_4 = 0,\ \lambda_2 = \frac{1}{2}, \lambda_1 + \lambda_3 = \frac{1}{2}) \text{ or} & \\ (\lambda_1 = 0,\ \lambda_3 = \frac{1}{2}, \lambda_2 + \lambda_4 = \frac{1}{2}) & \text{if } K = 4 \end{cases}.$$

A replica symmetric approximation

We say that the model is replica symmetric if the overlap is self averaging; what is the replica symmetric solution of a DBM model? This is important because in DL we know that the algorithms obtaining good classification performances are of RS type: belief propagation etc, and the local minima are wide, large entropy states.

The quenched pressure density of the model is now

$$p_{\Lambda_N}^{DBM}(\beta, h) \equiv \frac{1}{N}\,\mathbb{E}\log\sum_{\sigma}\exp\left(-\beta H_{\Lambda_N}(\sigma)+\sum_{p=1}^{K}\sum_{i\in L_p} h_i^{(p)}\sigma_i\right)$$

The replica symmetric approximation

Definition

Given $y = (y_p)_{p=1,\dots,K} \in [0,\infty)^K$ the *replica symmetric functional* is defined as

$$\mathcal{P}^{RS}_{\Lambda_N}(y,\beta,h) \equiv \sum_{p=1}^{K} \lambda_p^{(N)} \, \mathbb{E}_{z,h} \log\cosh\left(\beta\sqrt{2}\, q_p(\lambda,y)\, z + h^{(p)}\right)$$

$$+ \beta^2 \sum_{p=1}^{K-1} \lambda_p^{(N)} \lambda_{p+1}^{(N)} (1-y_p)(1-y_{p+1}) + \log 2$$

where $q_p(\lambda,y) = \sqrt{\lambda_{p-1}^{(N)}\, y_{p-1} + \lambda_{p+1}^{(N)}\, y_{p+1}}$ and
z is a standard Gaussian random variable independent of $h^{(1)},\dots,h^{(p)}$.

The replica symmetric approximation

The previous definition is motivated by the following *sum rule*:

$$p_{\Lambda_N}^{DBM}(\beta, h) = \mathcal{P}_{\Lambda_N}^{RS}(y, \beta, h) - \beta^2 \int_0^1 \left\langle R_N \right\rangle_{N,t} dt \ ,$$

where $\langle \cdot \rangle_{N,t}$ denotes the quenched Gibbs expectation associated to a suitable interpolating Hamiltonian and for every $\sigma, \tau \in \{\pm 1\}^N$

$$R_N(\sigma, \tau) \equiv \sum_{p=1}^{K-1} \lambda_p^{(N)} \lambda_{p+1}^{(N)} \left(q_{L_p}(\sigma, \tau) - y_p\right) \left(q_{L_{p+1}}(\sigma, \tau) - y_{p+1}\right) .$$

Stability condition for annealing

Stationary points of $\mathcal{P}^{RS}_{\Lambda_N}(y, \beta, h)$ satisfy the following system of self-consistent equations:

$$y_p \;=\; \mathbb{E}_z \, \tanh^2\left(\beta\sqrt{2}\sqrt{\lambda_{p-1}y_{p-1} + \lambda_{p+1}y_{p+1}}\; z + h_p\right) \qquad \forall\, p = 1, \ldots, K \;.$$

Now if we assume zero external field then $y = 0$ is a solution. Moreover

$$\mathcal{P}^{RS}(y=0, \beta, h=0, \lambda) \;=\; \log 2 \,+\, \beta^2 \sum_{p=1}^{K-1} \lambda_p \lambda_{p+1} \;.$$

Stability condition for annealing

A natural question is to ask for the conditions on β, λ that makes $y = 0$ stable

For a DBM with $K = 2, 3, 4$ layers one can prove that

The region of parameters (β, λ) such that the annealed solution $y = 0$ is stable coincide with the interior of the region A_K.

What's next?

- Need of interdisciplinary collaborations
- Need of interdisciplinary educational programs
- Need of an efficient communication to the public