January 2021, MAIn, PoliTo

A Mathematical-Physics Approach to Machine Learning

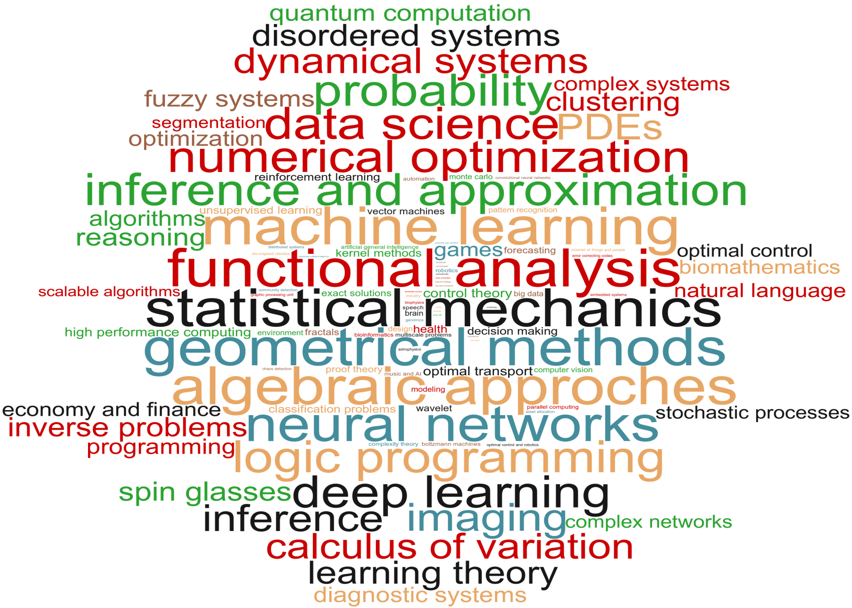
Pierluigi Contucci

Dipartimento di Matematica, Alma Mater Studiorum, Bologna

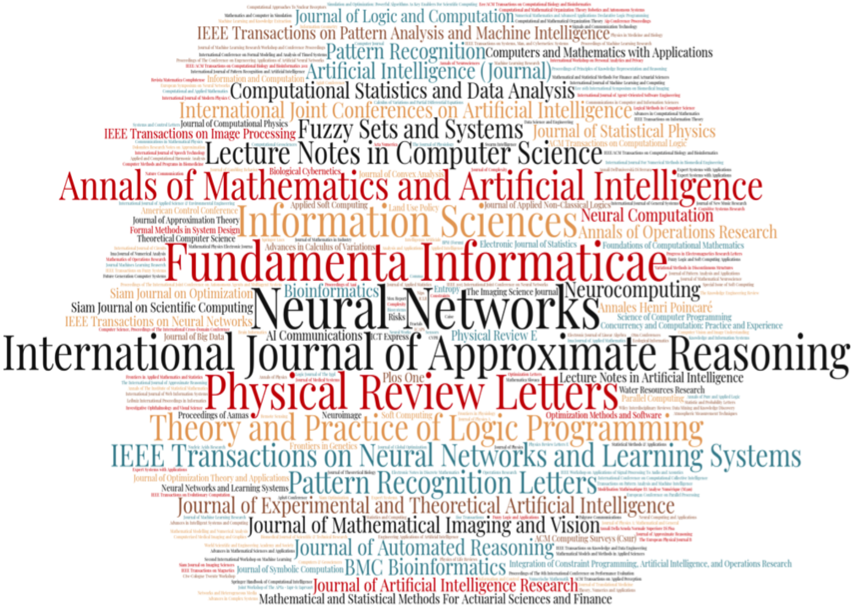
AI: how (why) does it work?

- AI and math, AI and the hard sciences, fantasy or **reality**?
- National poll: there are about 400 mathematicians, mostly already active in the field and some ready to step in.
- Mainly (but not only) in machine learning.
- The country (UMI, INDAM, the academic system) has the duty to support them.
- First step: awareness.

Keywords


quantum computation
disordered systems
dynamical systems
probability
complex systems
fuzzy systems
clustering
segmentation
data science
PDEs
optimization
numerical optimization
reinforcement learning
automation
monte carlo
inference and approximation
unsupervised learning
vector machines
pattern recognition
algorithms
machine learning
reasoning
kernel methods
games
forecasting
optimal control
functional analysis
biomathematics
scalable algorithms
exact solutions
control theory
big data
natural language
statistical mechanics
speech
brain
high performance computing
environment
fractals
design
health
decision making
bioinformatics
multiscale problems
geometrical methods
proof theory
music and AI
optimal transport
computer vision
algebraic approches
modeling
economy and finance
classification problems
wavelet
parallel computing
stochastic processes
inverse problems
neural networks
complexity theory
boltzmann machines
programming
logic programming
spin glasses
deep learning
inference
imaging
complex networks
calculus of variation
learning theory
diagnostic systems

Journals


Journal of Logic and Computation
IEEE Transactions on Pattern Analysis and Machine Intelligence
Pattern Recognition
Computers and Mathematics with Applications
Artificial Intelligence (Journal)
Information and Computation
Computational Statistics and Data Analysis
International Joint Conferences on Artificial Intelligence
Journal of Computational Physics
IEEE Transactions on Image Processing
Fuzzy Sets and Systems
Journal of Statistical Physics
ACM Transactions on Computational Logic
Lecture Notes in Computer Science
Biological Cybernetics
Journal of Convex Analysis
Annals of Mathematics and Artificial Intelligence
Applied Soft Computing
Land Use Policy
Journal of Applied Non-Classical Logics
American Control Conference
Journal of Approximation Theory
Formal Methods in System Design
Theoretical Computer Science
Information Sciences
Neural Computation
Annals of Operations Research
Advances in Calculus of Variations
Electronic Journal of Statistics
Foundations of Computational Mathematics
Fundamenta Informaticae
Siam Journal on Optimization
Bioinformatics
Entropy
The Imaging Science Journal
Neurocomputing
Siam Journal on Scientific Computing
Neural Networks
Risks
Annales Henri Poincaré
Science of Computer Programming
IEEE Transactions on Neural Networks
Concurrency and Computation: Practice and Experience
Journal of Big Data
AI Communications
ICT Express
Physical Review E
International Journal of Approximate Reasoning
Plos One
Lecture Notes in Artificial Intelligence
Physical Review Letters
Water Resources Research
Parallel Computing
Proceedings of Aamas
Neuroimage
Soft Computing
Optimization Methods and Software
Theory and Practice of Logic Programming
Frontiers in Genetics
IEEE Transactions on Neural Networks and Learning Systems
Journal of Optimization Theory and Applications
Neural Networks and Learning Systems
Pattern Recognition Letters
Journal of Experimental and Theoretical Artificial Intelligence
Journal of Mathematical Imaging and Vision
ACM Computing Surveys (Csur)
Journal of Automated Reasoning
Journal of Symbolic Computation
BMC Bioinformatics
Integration of Constraint Programming, Artificial Intelligence, and Operations Research
Journal of Artificial Intelligence Research
Mathematical and Statistical Methods For Actuarial Sciences and Finance

Machine Learning:

Machine

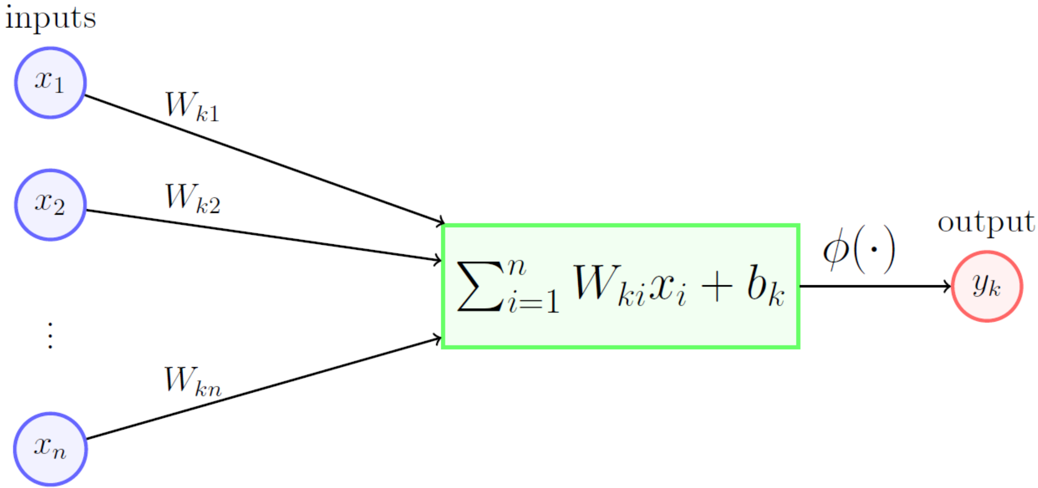
Input  **Output**

 **cat**

 **dog**

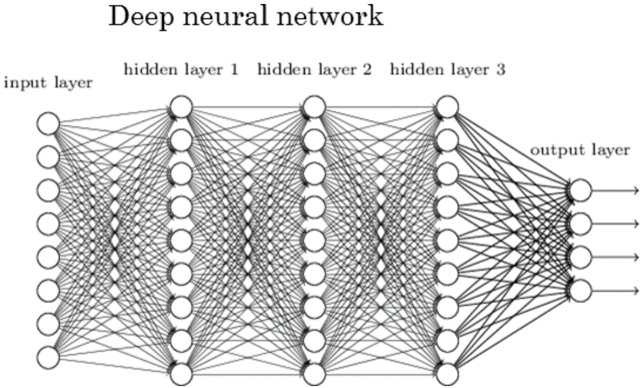
Adjust the parameters, Generalization ability!

Perceptron algorithm, Rosenblatt 1958



Bits of history

- Didn't work well, logical and practical objections (but Hinton persisted!)
- Improvements: shallow networks, deep networks



- Large databases, GPU, improved transmission and inversions algorithms, convolutional methods etc.

Efficacy and perspectives

- Nowadays works well in image recognition, language analysis, consumer profiling, and game playing (chess, go, etc).
- **Why? Are there principles and laws? Theory and models? Can we make predictions on if it's going to work on different environments? How about efficiency?**
- Deep Learning: technology, almost entirely heuristic, like pre-thermodynamic heat-engines.

Open problem (personal) collection

- Why is it convenient to grow in depth instead of wideness?
- Why deep networks do not overfit?
- Can we estimate the optimal number of parameters?
- What are the optimal form factors?
- Why good quality minima are easily found? Large entropy.

Open problem (personal) collection

Three classes of intertwined problems:

- **Data structure**: statistics, signal analysis
- **Modelisation**: physics, mathematics, math-phys
- **Algorithms**: computer science (beyond worse case), optimization

Boltzmann Machines

Math: what is the probabilistic model whose inferential solution is reached with machine learning? Boltzmann Machine!

- Precursors: Sherrington Kirkpatrick 1975, Parisi solution 1979, Hopfield Model 1982
- 1983, Hinton, Sejnowski
- 2009, Hinton, Salakhutdinov

Deep Learning: inverse problem, inference problem, with sampling assigned only in the boundary.

Paramount research: properties of the direct problem under simplifying hypoteses.

Mathematical setting

Let $V_N = \{1, \ldots, N\}$ be a set of labels for the N particles (or elementary agents, neurons) of a system.

- *Spin*: to each $i \in V$ we attach a binary variable $\sigma_i \in \{-1, 1\}$ representing the degree of freedom of a single particle.
- *Configuration space*: $\Sigma_N = \{-1, 1\}^N$. A point $\sigma = (\sigma_i)_{i \in V} \in \Sigma_N$ represents a (microscopic) configuration of the system.
- *Hamiltonian* or *energy* is a (random) function $H_N : \Sigma_N \to \mathbb{R}$

$$H_N(\sigma) = -\sum_{r,s \in \mathcal{S}} \sum_{\substack{i \in V_r, \\ j \in V_s}} W_{ij}^{(rs)} \sigma_i \sigma_j - \sum_{r \in \mathcal{S}} \sum_{i \in V_r} b_i^{(r)} \sigma_i \qquad (1)$$

Mathematical Setting

Study the properties of

- *Gibbs measure*: is a (random) measure on the configuration space Σ_N defined as

$$\mathcal{G}_N(\sigma) = \frac{1}{Z_N} e^{-\beta H_N(\sigma)} \tag{2}$$

where $Z_N = \sum_{\sigma \in \Sigma_N} e^{-\beta H_N(\sigma)}$ is the normalization or *partition function*.

- how does $\mathcal{G}_N$ behaves when $N \to \infty$ (thermodynamic limit)?
- compute the moments generating function in the thermodynamic limit

$$p_N = \frac{1}{N} \log Z_N \tag{3}$$

Multi-species disordered models

- Simplifying assumption: particles can be divided in different species (like in deep learning): let $\mathcal{S}$ be a finite set of labels with $|\mathcal{S}| = K$, we assume the vertex set V_N can be written as a disjoint union $V_N = \bigcup_{s \in \mathcal{S}} V_s$
- *Relative densities*: we assume that $\frac{|V_s|}{N} \to \alpha_s \in (0,1)$ for $N \to \infty$, for each $s \in \mathcal{S}$
- for $r, s \in \mathcal{S}$

$$W_{ij}^{(rs)} \overset{\text{iid}}{\sim} \mathcal{N}\left(\frac{\mu_{rs}}{2N}, \frac{\Delta_{rs}}{2N}\right), \tag{4}$$

$$b_i^{(s)} \overset{\text{iid}}{\sim} \mathcal{N}(\tilde{\mu}_s, \tilde{\Delta}_s) \tag{5}$$

- invariance under the direct product of the symmetric groups of each specie
- Key result: self averaging Gaussian concentration implies that

$$\lim_{N \to \infty} (p_N - \mathbb{E} p_N) = 0, \ W, b - \ a.s. \tag{6}$$

The convex multi-species model

Consider a finite set (of species) $\mathcal{S}$ and the Hamiltonian (1) assuming that:

- $\mu = 0$, centered interactions
- Δ is a semi-positive definite matrix
- $\tilde{\mu}$ and $\tilde{\Delta}$ are arbitrary

The property $\Delta \geq 0$ allows to extend the Parisi formula to the multi-species case, namely to express the quenched pressure for $N \to \infty$ as a (infinite dimensional) variational problem.

The core of the problem is the control interacting part of the Hamiltonian

$$H_N^{int}(\sigma) = -\sum_{r,s\in\mathcal{S}} \sum_{\substack{i\in V_r,\\ j\in V_s}} J_{ij}^{(rs)}\,\sigma_i\sigma_j \tag{7}$$

This is a centered Gaussian process with covariance

$$\mathbb{E}H_N^{int}(\sigma^1)H_N^{int}(\sigma^2) = N\sum_{r,s\in\mathcal{S}} \Delta_{rs}\,\alpha_r\,\alpha_s\; q_{12}^{(r)}q_{12}^{(s)} \tag{8}$$

where, for each $s \in \mathcal{S}$, we define the *relative overlap*

$$q_{12}^{(s)} = \frac{1}{|V_s|}\sum_{i\in V_s} \sigma_i^1\sigma_i^2 \tag{9}$$

Theorem (Barra-Contucci-Mingione-Tantari '13, Panchenko '13)

If the matrix Δ is positive semi-definite

$$\lim_{N\to\infty} p_N = \lim_{N\to\infty} \mathbb{E} p_N = \inf_x \mathcal{P}(x), \; J - \; a.s.$$

where $\mathcal{P}(x)$ is K-dimensional functional (generalization of Parisi functional for SK) and x is a cumulative distribution function on $[0,1]^K$ with suitable properties .

Ideas of the proof: The assumption $\Delta \geq 0$ allows to obtain an upper bound for $\mathbb{E}p_N$ dominating the interaction term by a suitable one body system (*replica symmetry breaking interpolation*). The converse bound is obtained exploiting the *synchronization property* of the overlap vector.

Deep Boltzmann machine (DBM)

In the Hamiltonian (1) with K species and centred interactions, consider the species arranged along a linear chain. Only pairs of consecutive species interact, while intra-species interactions as well as long range ones are forbidden. This amounts to the following assumptions on the parameters:

- Δ is a non-definite matrix, with zero diagonal and a tridiagonal structure

$$\Delta = \begin{pmatrix} 0 & \Delta_{12} & 0 & \cdots & 0 \\ \Delta_{12} & 0 & \Delta_{23} & \cdots & 0 \\ 0 & \Delta_{23} & 0 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \Delta_{K-1,K} \\ 0 & 0 & 0 & \Delta_{K-1,K} & 0 \end{pmatrix} ;$$

- $\mu = 0$, for centered interactions;
- $\tilde{\mu} = \tilde{\Delta} = 0$, in absence of external field.

Theorem (Alberici, Barra, Contucci, Mingione '20)

$$\liminf_{N\to\infty} \mathbb{E}\, p_N \geq \sup_a \mathcal{P}(\theta(a)) \ ,$$

where:

$$\mathcal{P}(\theta_1,\ldots,\theta_K) \;=\; \sum_{r=1}^{K} \Big(p^{\text{SK}}(\theta_r) - p^{\text{Ann-SK}}(\theta_r) \Big) \;+\; p^{\text{Ann.}} \ .$$

$p^{\text{SK}}(\theta_r)$ *denotes the limiting quenched pressure of a standard SK model at inverse temperature* θ_r *, while* $p^{\text{Ann-SK}}$ *denotes its annealed version and* $p^{\text{Ann.}}$ *the annealed pressure of the deep Boltzmann machine. Moreover*

$$\theta_r(a) \;=\; \sqrt{\alpha_r}\, \sqrt{\frac{1}{a_{r-1}}\, \Delta_{r-1,r} \;+\; a_r\, \Delta_{r,r+1}}$$

and the supremum is taken over $a = (a_1,\ldots,a_{K-1}) \in (0,\infty)^{K-1}$.

The annealed pressure is defined as

$$p^{\text{Ann.}} = \lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}\, Z_N$$

and is easy to compute thanks to the Gaussian nature of the interactions. Our lower bound provides the following result on the annealed regime of the deep Boltzmann machine:

Theorem (Alberici, Barra, Contucci, Mingione - Alberici, Contucci, Mingione '20)

If the tridiagonal matrix $M = \Delta \cdot \text{diag}(\alpha_1, \dots, \alpha_K)$ has spectral radius ≤ 1, then there exists

$$\lim_{N\to\infty} \mathbb{E}\, p_N = p^{\text{Ann.}} ,$$

namely the deep Boltzmann machine is in the annealed regime.

This result relies on the beautiful algebraic properties of certain Heilmann-Lieb polynomials, which establish a connection between deep Boltzmann machines and monomer-dimer systems.

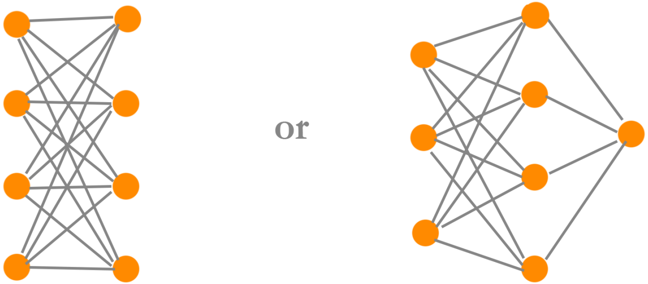
Minimizing the size of the annealed region for given interaction strengths, one finds the following optimal shapes for the DBM:

Theorem (Alberici, Barra, Contucci, Mingione - Alberici, Contucci, Mingione '20)

The maximum of the spectral radius of M *over* $\alpha_1, \ldots, \alpha_K \geq 0$, $\sum_r \alpha_r = 1$, *equals* $\max_r \Delta_{r,r+1}$ *and is reached if and only if:*

$$\alpha_{r^*} = \alpha_{r^*+1} = \frac{1}{2} \quad \textit{for } r^* \in \arg\max_r \Delta_{r,r+1}\,, \quad \textit{or:}$$

$$\alpha_{r^*-1} + \alpha_{r^*+1} = \alpha_{r^*} = \frac{1}{2} \quad \textit{for } r^*,\ r^*-1 \in \arg\max_r \Delta_{r,r+1}\,.$$



M-SK model on the Nishimori line

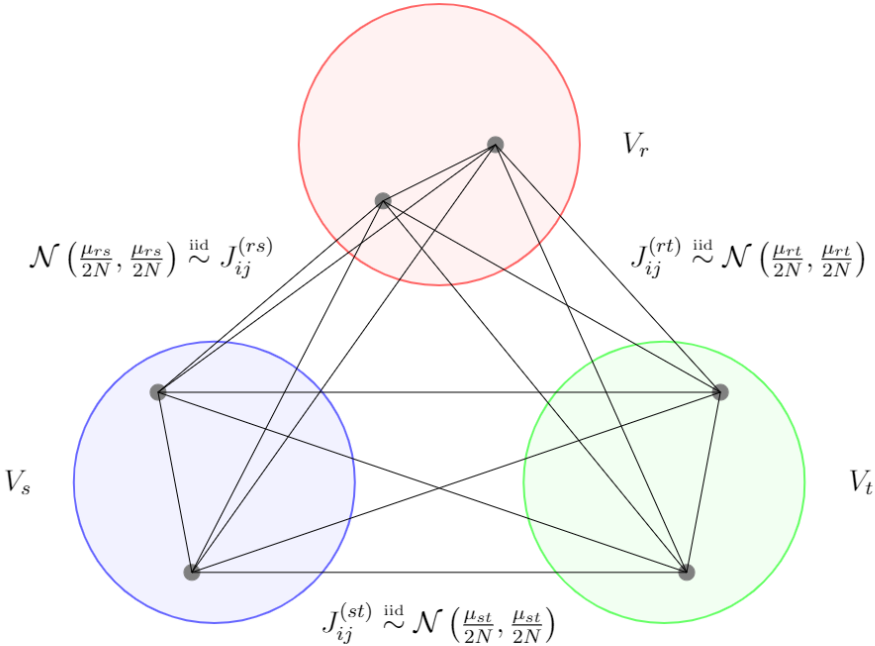
Nishimori line: a subregion of the space of parameters $\mu_{rs}, \Delta_{rs}, \tilde{\mu}_s, \tilde{\Delta}_s$ where $\mu_{rs} = \beta\Delta_{rs}$ and $\tilde{\mu}_s = \beta\tilde{\Delta}_s$. Reabsorbing β it is equivalent to have

$$J_{ij}^{(rs)} \overset{\text{iid}}{\sim} \mathcal{N}\left(\frac{\mu_{rs}}{2N}, \frac{\mu_{rs}}{2N}\right), \quad h_i^{(s)} \overset{\text{iid}}{\sim} \mathcal{N}(\tilde{\mu}_s, \tilde{\mu}_s).$$

Identities and correlation inequalities: models in this setting enjoy useful identities ($\langle\cdot\rangle_N$ denoting Boltzmann-Gibbs average) and inequalities (Contucci, Morita, Nishimori '04,'05):

$$\mathbb{E}[\langle\sigma_i\rangle_N^2] = \mathbb{E}[\langle\sigma_i\rangle_N], \quad \mathbb{E}[\langle\sigma_i\sigma_j\rangle_N^2] = \mathbb{E}[\langle\sigma_i\sigma_j\rangle_N],$$
$$\frac{\partial\mathbb{E}p_N}{\partial\mu_{rs}}, \frac{\partial\mathbb{E}p_N}{\partial\tilde{\mu}_s}, \frac{\partial^2\mathbb{E}p_N}{\partial\mu_{rs}^2}, \frac{\partial^2\mathbb{E}p_N}{\partial\tilde{\mu}_s^2} \geq 0$$

Replica symmetry: The previous ones imply replica symmetry in the models presented here. The variational principle is indeed finite dimensional.


V_r
$\mathcal{N}\left(\frac{\mu_{rs}}{2N}, \frac{\mu_{rs}}{2N}\right) \overset{\text{iid}}{\sim} J_{ij}^{(rs)}$
$J_{ij}^{(rt)} \overset{\text{iid}}{\sim} \mathcal{N}\left(\frac{\mu_{rt}}{2N}, \frac{\mu_{rt}}{2N}\right)$
V_s
V_t
$J_{ij}^{(st)} \overset{\text{iid}}{\sim} \mathcal{N}\left(\frac{\mu_{st}}{2N}, \frac{\mu_{st}}{2N}\right)$

The convex case on the Nishimori line

The thermodynamic limit is computed by means of the *adaptive interpolation method* (Barbier, Macris '19).

Theorem (Alberici, Camilli, Contucci, Mingione '20)

If μ is positive semidefinite, the thermodynamic limit of the random pressure converges J-a.s. and:

$$\lim_{N\to\infty} p_N = \lim_{N\to\infty} \mathbb{E} p_N = \sup_{\mathrm{x}\in[0,1]^K} \bar{p}(\mu, \tilde{\mu}; \mathrm{x})$$

where $\bar{p}(\mu, h; \mathrm{x})$ is a function of K parameters x.

When $\tilde{\mu}_s = 0$ the transition of x towards positive values is controlled by the spectral radius of $M := (\mu_{rs}\alpha_s)_{r,s=1,\dots,K}$.

- $\rho(M) < 1$: $\bar{p}$ is concave, $\mathrm{x} = 0$ is the unique maximizer;
- $\rho(M) > 1$: $\mathrm{x} = 0$ becomes an unstable saddle point.

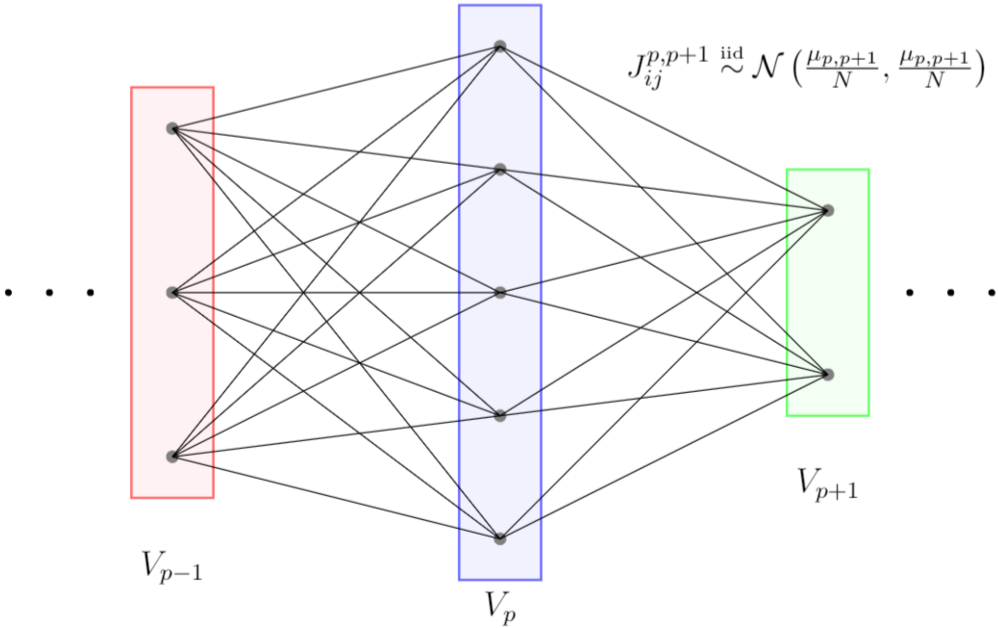
The deep Boltzmann machine on the Nishimori line

The DBM on the Nishimori line corresponds to the choice:

$$\mu = \begin{pmatrix} 0 & \mu_{12} & 0 & \cdots & 0 \\ \mu_{21} & 0 & \mu_{23} & \cdots & 0 \\ 0 & \mu_{32} & 0 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \mu_{K-1,K} \\ 0 & 0 & 0 & \mu_{K,K-1} & 0 \end{pmatrix},$$

namely, species are arranged in a consecutive way and only adjacent species are allowed to interact. Intra-group interactions are forbidden (0 diagonal elements).

μ has eigenvalues with alternating sign (symmetric w.r.t. 0). This clearly violates the positivity hypothesis of the convex case.


$J_{ij}^{p,p+1} \overset{\text{iid}}{\sim} \mathcal{N}\left(\frac{\mu_{p,p+1}}{N}, \frac{\mu_{p,p+1}}{N}\right)$
V_{p-1}
V_p
V_{p+1}

The thermodynamic limit of the DBM

Theorem (Alberici, Camilli, Contucci, Mingione '20)

The random pressure of the deep Boltzmann machine on the Nishimori line converges J-a.s. and

$$\lim_{N\to\infty} p_N = \lim_{N\to\infty} \mathbb{E} p_N = \sup_{\mathrm{x}_o} \inf_{\mathrm{x}_e} p_{var}(\mu, \tilde{\mu}; \mathrm{x}) \ ,$$

where x_o *and* x_e *denote the vectors of the odd and even components of* $\mathrm{x} \in [0,1)^K$ *respectively.*

When $\tilde{\mu}_s = 0$ the transition of x towards positive values is controlled by the submatrix (recall $M_{rs} = \mu_{rs}\alpha_s$) $[M^2]^{(oo)} := ((M^2)_{rs})_{r,s \text{ odd } \leq K}$:

- $\rho([M^2]^{(oo)}) < 1$: $\mathrm{x} = 0$ is the unique optimizer;
- $\rho([M^2]^{(oo)}) > 1$: the optimizer $\mathrm{x} = \bar{\mathrm{x}}$ has striclty positve components.

Bibliography, Math.

- On the equivalence of Hopfield Networks and Restricted Boltzmann Machines (A.Barra, A.Bernacchia, P.C., E.Santucci) Neural Networks (2012)
- Multi-species mean-field spin-glasses. Rigorous results (A.Barra, P.C. E.Mingione and D.Tantari) Annales Henri Poincare, (2014)
- A Multi-Scale Sherrington-Kirkpatrick Model. (P.C., E.Mingione) Communications in Mathematical Physics, (2019)
- Annealing and replica-symmetry in Deep Boltzmann Machines (D.Alberici, A.Barra, P.C. E.Mingione) Journal of Statistical Physics (2020)
- The multi-species mean-field spin-glass in the Nishimori line (D. Alberici, F. Camilli, P.C., E.Mingione) Journal of Statistical Physics (2020)
- Deep Boltzmann machines: rigorous results at arbitrary depth (D.Alberici, P.C. E.Mingione) To appear in Annals Henri Poincare (2020)
- The solution of the deep Boltzmann machine on the Nishimori line (D. Alberici, F. Camilli, P.C., E. Mingione) Preprint (2020)

Bibliography, General

- Intelligenza artificiale tra rischi e opportunità Alcune riflessioni sul tema. P.C. Rivista Il Mulino, N. 4, 637-645, (2019)

- Fisici e Matematici in Intelligenza Artificiale. (in collaboration with P.Branchini) Notiziario dell'Unione Matematica Italiana, settembre (2019)

- Intelligenza artificiale e apprendimento automatico. P.C. Rivista il Mulino Online, ottobre (2018)

- AI, a resource for Italy. Strategies for the future. P.C. Webinar transcript, http://contucci.blogspot.com/2020/09/estratto-del-mio-contributo-al-webinar.html