La macchina di Boltzmann: quando il modello di Ising incontra il Machine Learning

Aurélien Decelle

Departamento de Física Téorica I, Universidad Complutense, 28040 Madrid, Spain,
TAU, LRI, INRIA, Université Paris Sud, CNRS, Université Paris Saclay, Orsay 91405, France.

Il machine learning sta diventando sempre più importante nella ricerca e nella vita quotidiana, tuttavia il processo dell'apprendimento rimane in gran parte oscuro e molte questioni importanti sono ancora irrisolte. I meccanici statistici, in una lunga tradizione di ricerca di comportamenti universali e meccanismi semplici per comprendere fenomeni collettivi complessi, hanno provato a comprendere questi modelli con il loro linguaggio. È quindi naturale che la macchina di Boltzmann - o il problema inverso di Ising - si inserisca nell'intersezione tra meccanica statistica e machine learning.

Machine Learning is becoming more and more important in research and in daily life, yet the learning process remains largely misunderstood and many important questions are still unresolved. Statistical physicists, in a long tradition of looking for universal behavior and simple mechanisms to understand complex collective phenomena, have taken their turn in trying to understand these models with their own language. It is therefore natural that the Boltzmann Machine - or the inverse Ising problem - enters at the crossroad of statistical physics and machine learning.

Introduzione

È inutile ricordare al lettore l'importanza che ha assunto recentemente il Machine Learning (ML) nella nostra vita quotidiana [1]. Eppure, anche con le promesse della GAFAM (Google, Amazon, Facebook, Apple, Microsoft, ...) di utilizzare l'intelligenza ar-

Introduction

It is useless to remind the reader the importance that has taken recently Machine Learning (ML) in our daily lives[1]. Yet, even with the promises of the GAFAM (Google, Amazon, Facebook, Apple, Microsoft, ...) to bring use artifical intelligence to im-

tificiale per migliorare la nostra vita quotidiana, la ricerca in questo campo fatica ancora a spiegare il perché e come i modelli di machine learning funzionano nei dettagli. Tuttavia, i progressi nel ML sono stati enormi negli ultimi decenni. Non molto tempo fa, era difficile e macchinoso eseguire un compito di classificazione "semplice" su un insieme di immagini (inferendo automaticamente una categoria di un oggetto in un'immagine per esempio) mentre oggigiorno con strumenti moderni può essere fatto da chiunque in grado di scrivere codice in Python, un popolare linguaggio di programmazione. Nonostante questi progressi, la comprensione dei meccanismi fondamentali che avvengono nel processo di apprendimento rimane ampiamente elusa.

La meccanica statistica (ma non solo) ha una lunga tradizione nel cercare di comprendere problemi che esulano dal suo campo "tipico". Ad esempio, negli anni '80 e '90, molti fisici (ed ovviamente matematici) iniziarono a studiare modelli dell'informatica, come problemi di ottimizzazione e reti neurali. È quindi naturale che il recente sviluppo del ML abbia rinnovato l'interesse dei fisici per questo campo, specialmente di fronte all'enorme successo di queste macchine.

In questo contributo ci concentreremo su una particolare tipologia di macchina, introdotta molti decenni fa: la Boltzmann Machine (BM), e più precisamente sulla sua versione "ristretta". La BM è stata introdotta da Hinton e Sejnowksi [2] come una "constraint satisfaction network [...] capable of learning the underlying constraint". L'idea è di modificare l'intensità delle connessioni della rete per costruire un modello generativo interno che produca campioni secondo la stessa distribuzione di probabilità del set di dati fornito. La definizione della BM seguirà nella prossima sezione, ma vorremmo sottolineare qui che questo modello include già l'ingrediente interessante per i fisici: la BM corrisponde al problema di Ising inverso. Basandoci su un insieme di campioni — o configurazioni di Ising — l'obiettivo è ricostruire la rete, ovvero inferire le costanti di accoppiamento tra gli spin di Ising. Successivamente, Hinton [3] ha introdotto la sua versione ristretta, in cui viene usata una struttura bipartita tra le variabili, aprendo la possibilità di abbinare statistiche di ordine superiore tra la distribuzione dedotta (dalla macchina) ed il set di dati forniti (alla macchina). Gli aspetti interessanti di questi modelli per il fisico sono che, in primo luogo, il processo di apprendimento corrisponde alla

prove our daily life, the research in this field is still struggling to explain the why and how the machine learning models work in details. Still, the progress in ML has performed a huge advances in the last decades. Not so long ago, it was hard and cumbersome to perform a "simple" classification task on a set of images (automatically inferring a category of an object in an image for instance) whereas nowadays with modern tools it can be done by anyone capable of coding in Python, a popular programming language. Despite these progresses, the understanding of fundamental mechanisms taking place in the learning process remains largely misunderstood.

Statistical mechanics (but not only) has a long tradition of trying to understand problems that lie outside its "typical" field. For instance, in the 80' and 90', many physicists (and mathematicians obviously) began to study models from computer science, such as optimization problems and neural networks. It is therefore only natural that the recent development in ML renewed the interest of physicists for this field, specially in front of the huge success of these machines.

In this contribution, we will focus on a particular type of machine, which has been introduced many decades ago: the Boltzmann Machine (BM), and more specifically on its "restricted" version. The BM has been introduced by Hinton and Sejnowksi[2] as a "constraint satisfaction network [...] capable of learning the underlying constraint". The idea is to modify the strength of the network connections to construct an internal generative model that produces samples following the same probability distribution as a provided dataset. The definition of the BM will follow in the next section but we would like to emphasis here that this model already includes the interesting ingredient for physicists: the BM corresponds to the inverse Ising problem: based on a set of samples —or Ising configurations— the goal is to "reconstruct the network", namely to infer coupling constant between the Ising spins. Later on, Hinton[3] introduced its restricted version, where a bipartite structure between the variables is introduced, opening the possibility to match higher order statistics between the inferred distribution and the dataset. The interesting aspects of these models for physicist are that, first the learning process corresponds to the inverse procedure of the disordered Ising model, a model that has been studied

procedura inversa del modello di Ising disordinato, un modello che è stato studiato per più di un secolo. Secondo, le fasi di equilibrio di una macchina in cui i parametri sono stati appresi su un dataset non banale devono ancora essere esplorate.

Nella nostra prossima analisi, inizieremo definendo il modello di Ising e mostreremo come il teorema di Bayes fornisca una struttura naturale per procedere con il problema di Ising inverso, o equivalentemente con la BM. Vedremo poi come costruire la Restricted Boltzmann Machine (RBM), partendo da una semplice descrizione utilizzando la rete bipartita, e come, rendendo il modello più complesso, arriveremo naturalmente alla macchina descritta da Hinton, e capace di modellazione di set di dati complessi.

Il modello di Ising

Il modello di Ising è un oggetto ben noto per i meccanici statistici. La definizione è molto semplice, prendendo un insieme di N spin di Ising: $s_i = \pm 1$, definiamo la seguente Hamiltoniana

$$\mathscr{H}[\underline{s}] = -\sum_{i<j} J_{ij} s_i s_j - \sum_i h_i s_i \qquad (1)$$

dove J_{ij} sono le costanti di accoppiamento tra gli spin. La distribuzione di Boltzmann è quindi data da

$$p_{Ising}(\boldsymbol{s}) = \frac{1}{Z} e^{\beta \mathscr{H}[\underline{s}]}, \text{ con } Z = \sum_{\{s\}} e^{\beta \mathscr{H}[\underline{s}]} \qquad (2)$$

Z è la funzione di partizione, $\beta = 1/T$ la temperatura inversa, e indichiamo $\langle f(\boldsymbol{s}) \rangle_{\mathscr{H}}$ la media termica, effettuata rispetto alla probabilità data dall'eq. (2). Il fenomeno classico descritto da questo sistema è il ferromagnetismo: quando $J_{ij} = 1$ il sistema appare con due fasi distinte. Una fase paramagnetica, dove la media $m_i = \langle s_i \rangle = 0$, ad alta temperatura ed una fase ferromagnetica, che appare improvvisamente a $\beta = \beta_c$, dove il sistema mostra una magnetizzazione spontanea $m_i \neq 0$.

L'approccio standard per studiare il modello di Ising, a seconda della sua struttura di accoppiamenti J_{ij}, consiste nel calcolare l'energia libera del sistema $f = -N^{-1} \log(Z)$ per stabilirne il diagramma di fase: ciò permette di osservarne il comportamento macroscopico al variare dei suoi parametri. Nel caso del ferromagnete, il parametro di controllo è β mentre il comportamento macroscopico si deduce dal

for more than a century now. Second the equilibrium phases of a machine where the parameters have been learned on a non trivial dataset is still to be explored.

In our upcoming analysis, we will start by defining the Ising model and how Bayes theorem provides a natural framework to proceed with the inverse Ising problem, or equivalently the BM. Then, we will see how we can construct the Restricted Boltzmann Machine (RBM), starting from a simple description using the bipartite network, and how, making the model more complex, we will naturally arrive to the machine described by Hinton, and capable of modeling complex dataset.

The Ising model

The Ising model is a well-known object for statistical physicists. The definition is very simple, taking a set of N Ising spins: $s_i = \pm 1$, we define the following Hamiltonian

$$\mathscr{H}[\underline{s}] = -\sum_{i<j} J_{ij} s_i s_j - \sum_i h_i s_i \qquad (1)$$

where the J_{ij} are the coupling constants between the spins and the h_i the local magnetic fields. The Boltzmann distribution is then given by

$$p_{Ising}(\boldsymbol{s}) = \frac{1}{Z} e^{\beta \mathscr{H}[\underline{s}]}, \text{ with } Z = \sum_{\{s\}} e^{\beta \mathscr{H}[\underline{s}]} \qquad (2)$$

Z being the partition function, $\beta = 1/T$ the inverse temperature, and we will denote $\langle f(\boldsymbol{s}) \rangle_{\mathscr{H}}$ the thermal average with respect to the probability given by eq. (2). The classical phenomena described by this system is the ferro-magnetism, where $J_{ij} = 1$ and $h_i = 0$ describing a system with two distinct phases: a paramagnetic phase where the average $m_i = \langle s_i \rangle = 0$ at high temperature, and a ferromagnetic phase appearing suddenly at $\beta = \beta_c$ where the system shows a spontaneous magnetization $m_i \neq 0$.

The standard approach to study the Ising model, depending of its couplings structure J_{ij} and local fields h_i, consists in computing the free energy of the system $f = -N^{-1} \log(Z)$ in order to establish the phase diagram of the system, exhibiting its macroscopical behavior as a function its parameters. In the case of the ferromagnet, the control parameter is β and the macroscopic behavior defined by the value of the

valore della magnetizzazione m_i, cioè del parametro d'ordine.

Possiamo ora definire il problema di Ising inverso che mira a rispondere alla seguente domanda: avendo un set di configurazioni di spin $\{\boldsymbol{s}^{(d)}\}$ dove $d = 1,\dots,M$, possiamo identificare un set di parametri $\theta = \{J,h\}$ per cui la distribuzione Boltzmann ad essi associata riproduce, statisticamente, le stesse configurazioni? Utilizzando il teorema di Bayes, è possibile definire la seguente probabilità sui parametri di Ising

$$p(\theta|\boldsymbol{s}^{(d)}) \propto \prod_d^M \left[p_{\mathrm{I}sing}(\boldsymbol{s}^{(d)}|\theta)\right] p_{\mathrm{p}rior}(\theta) \quad (3)$$

dove $p_{\mathrm{p}rior}$ è una distribuzione a priori, i.e. una prior, sui parametri da inferire (aggiungiamo la dipendenza dei parametri nella probabilità di Boltzmann). Senza discutere sui molti possibili stimatori per i parametri θ, senza la prior, possiamo vedere dall'eq. (3) che il massimo del membro di sinistra può essere ottenuto massimizzando la verosimiglianza (o log-verosimiglianza) del modello

$$\mathscr{L} = \left[\beta \sum_{i<j} J_{ij}\langle s_i s_j\rangle_{\mathrm{d}} + \sum_i h_i \langle s_i\rangle_{\mathrm{d}}\right] - \log Z \quad (4)$$

dove $\langle f(\boldsymbol{s})\rangle_{\mathrm{d}} = \frac{1}{M}\sum_{d=1}^M f(\boldsymbol{s}^{(d)})$ è la media di una funzione f rispetto al set di dati. È impossibile massimizzare direttamente l'eq. (4) a causa della funzione di partizione, che è computazionalmente intrattabile in generale. Tuttavia, è possibile calcolare approssimativamente i gradienti rispetto alle J_{ij} ed ai h_i: questi sono dati da

$$\frac{\partial \mathscr{L}}{\partial J_{ij}} = \beta\left(\langle s_i s_j\rangle_{\mathrm{d}} - \langle s_i s_j\rangle_{\mathscr{H}}\right)$$
$$\frac{\partial \mathscr{L}}{\partial h_i} = \beta\left(\langle s_i\rangle_{\mathrm{d}} - \langle s_i\rangle_{\mathscr{H}}\right)$$

e permettono di eseguire una risalita del gradiente: aggiornando ad ogni iterazione t il valore degli accoppiamenti, utilizziamo il seguente schema iterativo

$$J_{ij}^{(t+1)} = J_{ij}^{(t)} + \eta \left.\frac{\partial \mathscr{L}}{\partial J_{ij}}\right|_{J_{ij}^{(t)},h_i^{(t)}}$$
$$h_i^{(t+1)} = h_i^{(t)} + \eta \left.\frac{\partial \mathscr{L}}{\partial h_i}\right|_{J_{ij}^{(t)},h_i^{(t)}}$$

dove η è chiamato tasso di apprendimento e regola la velocità con cui stiamo cambiando i parametri.

magnetization m_i, i.e. the order parameter.

We can now define the inverse Ising problem which aims to answer to the following question. Having a set of spins configuration $\{\boldsymbol{s}^{(d)}\}$ where $d = 1,\dots,M$, can we identify a set of parameters $\theta = \{J,h\}$ for which the associated Boltzmann distribution reproduced the same statistics ? Using the Bayes theorem, it is possible to defined the following probability over the Ising parameters

$$p(\theta|\boldsymbol{s}^{(d)}) \propto \prod_d^M \left[p_{\mathrm{I}sing}(\boldsymbol{s}^{(d)}|\theta)\right] p_{\mathrm{p}rior}(\theta) \quad (3)$$

where $p_{\mathrm{p}rior}$ is a prior distribution over the parameters to infer (we add the dependence of the parameters in the Boltzmann probability). Without debating on the many possible estimators for the parameters θ, in the absence of prior, we can see from eq. (3) that maximizing the l.h.s. can be achieved by maximizing the likelihood (or log-likelihood) of the model

$$\mathscr{L} = \left[\beta \sum_{i<j} J_{ij}\langle s_i s_j\rangle_{\mathrm{d}} + \sum_i h_i \langle s_i\rangle_{\mathrm{d}}\right] - \log Z \quad (4)$$

where $\langle f(\boldsymbol{s})\rangle_{\mathrm{d}} = \frac{1}{M}\sum_{d=1}^M f(\boldsymbol{s}^{(d)})$ is the average of a function f over the dataset. It is impossible to maximize directy eq. (4) because of the partition function which is intractable in general. However, it is easier to compute approximately the gradients over J_{ij} and h_i, which are given by

$$\frac{\partial \mathscr{L}}{\partial J_{ij}} = \beta\left(\langle s_i s_j\rangle_{\mathrm{d}} - \langle s_i s_j\rangle_{\mathscr{H}}\right)$$
$$\frac{\partial \mathscr{L}}{\partial h_i} = \beta\left(\langle s_i\rangle_{\mathrm{d}} - \langle s_i\rangle_{\mathscr{H}}\right)$$

and to perform a gradient ascent, updating at each iteration t the value of the couplings using the following iterative scheme

$$J_{ij}^{(t+1)} = J_{ij}^{(t)} + \eta \left.\frac{\partial \mathscr{L}}{\partial J_{ij}}\right|_{J_{ij}^{(t)},h_i^{(t)}}$$
$$h_i^{(t+1)} = h_i^{(t)} + \eta \left.\frac{\partial \mathscr{L}}{\partial h_i}\right|_{J_{ij}^{(t)},h_i^{(t)}}$$

where η is called the learning rate and adjusts the velocity with which we are changing the parameters. For this problem, it can be shown that the likelihood is convex and thus the convergence is guaranteed. However, the second term of the r.h.s. of the gradient is usually approximated using samples obtained from Monte Carlo Markov Chain (MCMC) simulations.

Per questo problema si può dimostrare che la verosimiglianza è convessa e quindi la convergenza è garantita. Tuttavia, il secondo termine dei membri di destra di questo gradiente viene solitamente approssimato utilizzando campioni ottenuti da simulazioni mediante catene di Markov Monte Carlo (MCMC). Di conseguenza, la convergenza verso il massimo può essere lenta a causa del tempo di mixing della catena e dell'errore statistico.

Infine, il problema inverso di Ising può anche essere derivato utilizzando il principio di massima entropia. Questa tecnica forza, mediante l'impiego dei moltiplicatori di Lagrange, a far corrispondere le statistiche di basso ordine di una distribuzione di probabilità –la magnetizzazione di ogni variabile $\langle s_i \rangle_{\mathscr{H}}$ e tutte le correlazioni di coppia $\langle s_i s_j \rangle_{\mathscr{H}}$– con quelle estrattte da un set di dati, imponendo che l'entropia della distribuzione ricavata sia massima. In questa formulazione i moltiplicatori di Lagrange vengono quindi identificati con i campi magnetici $\boldsymbol{h}$ e con la matrice di accoppiamenti $\boldsymbol{J}$: in concreto si impone che la magnetizzazione e tutte le correlazioni a coppie del set di dati corrispondano a quelle che la distribuzione ottenuta genera.

La macchina di Boltzmann

La BM corrisponde esattamente al problema inverso di Ising, ma BM è il termine più comunemente usato nell'informatica. Quando è stata introdotta per la prima volta, le variabili erano discrete con valori in $\{0,1\}$ invece di spin di Ising. Questa scelta cambia solo la parametrizzazione del modello poiché un semplice cambio di variabile collega le due formulazioni. Per i fisici, questa formulazione è alquanto innaturale poiché rompe la simmetria di spin-flip da $s_i \to -s_i$ presente nell'Hamiltoniana in assenza di campi magnetici. La ragione di questa scelta è che, usando $\{0,1\}$, una variabile può essere vista come *attiva*, quando $s_i = 1$, o *inattiva* quando $s_i = 0$. Questa terminologia si riferisce al fatto che una variabile attiva contribuirà al campo effettivo locale di una a lei connessa $s_i = \sum_{j \neq i} J_{ij} s_j + h_i$, mentre una inattiva no. Infine, la matrice di accoppiamento è anche chiamata matrice dei pesi (sinaptici) ed indicata con $\boldsymbol{w}$, mentre i campi magnetici sono chiamati bias.

Finora, la BM è una macchina identica a quella utilizzata nel problema di Ising inverso. Di conseguenza, la limitazione principale è la stessa: queste macchine sono fatte per regolare i loro parametri al

Consequently, the convergence to the maximum can be slow due to the mixing time of the chain and the statistical error.

Finally, the Ising inverse problem can also be derived using the maximum entropy principle. This technique aimed at matching the low-order statistics of a dataset —the magnetization of each variable $\langle s_i \rangle_{\mathscr{H}}$ and all the pairwise correlations $\langle s_i s_j \rangle_{\mathscr{H}}$— for a probability distribution using Lagrange multipliers, imposing that the target distribution's entropy is maximal. In this formulation, the Lagrange multipliers are then identified to the the magnetic fields $\boldsymbol{h}$ and the couplings matrix $\boldsymbol{J}$. They enforce the magnetization and all the pairwise correlations of the dataset to match the ones of the target distribution.

The Boltzmann Machine

The BM corresponds exactly to the inverse Ising problem, the term being more commonly used in computer science. When it was introduced for the first time, the variables were discrete with values in $\{0,1\}$ instead of Ising spins. This only change parametrization of the model since a simple change of variable links the two formulations. For physicists, this formulation is somewhat unnatural since it breaks the spin-flip symmetry $s_i \to -s_i$ present in the Hamiltonian in the absence of magnetic fields. The reason behind this choice is that, using $\{0,1\}$, a variable can be seen as "active", when $s_i = 1$, or "inactive" when $s_i = 0$. This terminology refers to the fact that an active variable will contribute to the local effective field of a neighbor $s_i = \sum_{j \neq i} J_{ij} s_j + h_i$, whereas an inactive will not. Finally, the coupling matrix is also named a (synaptic) weight matrix and denoted as $\boldsymbol{w}$, and the magnetic fields are called biases.

So far, the BM is a completely similar machine as the one used in the inverse Ising problem. Consequently, the major limitation is the same: these machines are made to adjust their parameters in order to match the magnetization and pairwise correlations

fine di abbinare la magnetizzazione e le correlazioni a coppie di un set di dati. Ma, nella loro formulazione classica, non possono regolare alcun parametro al fine di abbinare statistiche di ordine superiore.

of a dataset. But, in their classical formulation they cannot tune any parameters in order to match higher-order statistics.

Dalla BM alla RBM

È di nuovo Hinton a proporre l'idea della Restricted Boltzmann Machine [3]. La RBM è un'estensione della BM, dove viene introdotto un nuovo insieme di variabili $\{0,1\}$ discrete: le variabili nascoste (o latenti). Di seguito chiameremo i nodi visibili s_i, con $i = 1,\dots,N_v$ e i nodi nascosti τ_a, con $a = 1,\dots,N_h$. Oltre all'introduzione di nuove variabili, ora questi due insiemi di variabili risiedono in due strati differenti e le interazioni esistono solo tra nodi di diversi strati. come è illustrato nella Figura 1.

La matrice dei pesi, quindi, avrà una componente diversa da zero solo tra un nodo visibile e uno nascosto, definendo quindi l'Hamiltoniana successiva

$$\mathscr{H}[\boldsymbol{s},\boldsymbol{\tau}] = -\sum_{i,a=1}^{(N_v,N_h)} w_{ia}s_i\tau_a - \sum_i h_i s_i - \sum_a \bar{h}_a\tau_a$$

From BM to RBM

It is again Hinton who came with the idea of the Restricted Boltzmann Machine[3]. The RBM is an extension of the BM, where a new set of discrete $\{0,1\}$ variables is introduced: the hidden (or latent) variables. In the following we will call the visible nodes s_i, with $i = 1,\dots,N_v$ and the hidden nodes τ_a, with $a = 1,\dots,N_h$. In addition to the introduction of new variables, the two sets of variables lives in two different layers and interactions will exist only between nodes of different layers, as can be seen on Figure 1.

The weight matrix, will therefore have non-zero component only between a visible and a hidden node, defining the following Hamiltonian

$$\mathscr{H}[\boldsymbol{s},\boldsymbol{\tau}] = -\sum_{i,a=1}^{(N_v,N_h)} w_{ia}s_i\tau_a - \sum_i h_i s_i - \sum_a \bar{h}_a\tau_a$$

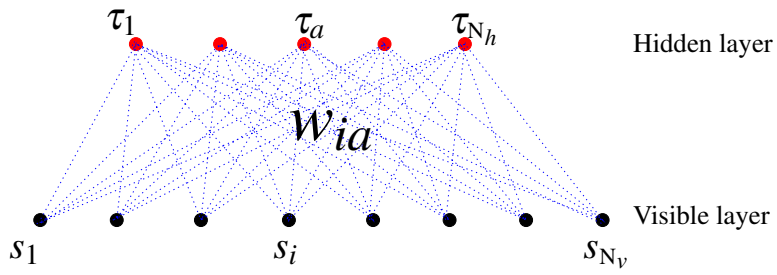

Figura 1: *Struttura bipartita della RBM.*
Bipartite structure of the RBM.

Il ruolo dei nodi nascosti è quello di tenere conto dell'effettiva interazione tra i nodi visibili. Queste interazioni avvengono quando si marginalizza sulle variabili nascoste, dando la seguente distribuzione di probabilità sui nodi visibili.

$$p(\boldsymbol{s}) = \sum_{\{\boldsymbol{\tau}\}} p(\boldsymbol{s},\boldsymbol{\tau}) = \frac{1}{Z}\sum_{\{\boldsymbol{\tau}\}} \exp\left(-\mathscr{H}[\boldsymbol{s},\boldsymbol{\tau}]\right)$$
$$= e^{\sum_i h_i s_i}\prod_a \left(1 + exp\left(\sum_i w_{ia}s_i + \bar{h}_a\right)\right)$$

La distribuzione dei nodi visibili mostra un insieme ricco e complesso di interazioni in cui eventualmente

The role of the hidden nodes is to take into account the effective interaction between the visible nodes. These interactions take place when marginalizing over the hidden variables, giving the following probability distribution over the visible nodes.

$$p(\boldsymbol{s}) = \sum_{\{\boldsymbol{\tau}\}} p(\boldsymbol{s},\boldsymbol{\tau}) = \frac{1}{Z}\sum_{\{\boldsymbol{\tau}\}} \exp\left(-\mathscr{H}[\boldsymbol{s},\boldsymbol{\tau}]\right)$$
$$= e^{\sum_i h_i s_i}\prod_a \left(1 + \exp\left(\sum_i w_{ia}s_i + \bar{h}_a\right)\right)$$

The distribution over the visible nodes exhibits a rich and complex set of interactions where possibly more

più di due nodi possono essere in interazione diretta, a seconda della matrice dei pesi. Infatti, facendo una semplice espansione in w_{ia} piccoli, possiamo ottenere un'Hamiltoniana efficace contenente interazioni in qualsiasi ordine tra i nodi visibili. Infatti, è stato dimostrato che le RBM sono approssimatori universali [4] di distribuzioni discrete, cioè una RBM opportunamente grande può approssimare arbitrariamente bene qualsiasi distribuzione discreta.

La procedura di apprendimento della RBM è simile a quella della BM, la differenza sta nel fatto che la distribuzione non appartiene pià alla famiglia esponenziale. Tuttavia, è ancora possibile ottenere una risalita del gradiente calcolando la derivata della verosimiglianza

$$\frac{\partial \mathscr{L}}{\partial w_{ia}} = \langle s_i \sum_{\tau_a=0,1} \tau_a p(\tau_a|\boldsymbol{s})\rangle_{\mathrm{d}} - \langle s_i \tau_a \rangle_{\mathscr{H}}$$
$$\frac{\partial \mathscr{L}}{\partial h_i} = \langle s_i \rangle_{\mathrm{d}} - \langle s_i \rangle_{\mathscr{H}}$$
$$\frac{\partial \mathscr{L}}{\partial \bar{h}_a} = \langle \sum_{\tau_a=0,1} \tau_a p(\tau_a|\boldsymbol{s})\rangle_{\mathrm{d}} - \langle \tau_a \rangle_{\mathscr{H}}$$

Una differenza notevole è la presenza della media condizionata sul nodo nascosto a nel termine mediato sul set di dati. In pratica, non introduce complicazioni concrete in quanto la distribuzione condizionata fattorizza sul valore dei nodi visibili

$$p(\tau_a = 1|\boldsymbol{s}) = \frac{1}{1+\exp\left(\sum_i w_{ia} s_i + \bar{h}_a\right)}$$

Di conseguenza, la procedura di addestramento per la RBM è simile a quella della BM. All'opposto della BM, la verosimiglianza della RBM non è più convessa rispetto alla media: i parametri del modello e quindi la salita del gradiente, in generale, non convergeranno al previo insieme di parametri. Un risultato interessante dell'architettura bipartita del modello è che, condizionata su uno strato (visibile o nascosto), la distribuzione marginale rispetto alle variabili dell'altro strato fattorizza. Utilizzando queste proprietà possiamo implementare un'efficiente procedura di campionamento: fissando il valore delle variabili di un dato strato, le variabili dell'altro strato possono essere estratte in parallelo molto velocemente.

La RBM è un oggetto molto complesso da analizzare, sia inerentemente le sue proprietà di equilibrio che il suo processo di apprendimento. Di seguito mostreremo innanzitutto come una RBM gaussiana, nonostante sia molto più semplice, mostri un com-

than two nodes can be in direct interaction, depending on the weight matrix. In fact, making a simple expansion in small w_{ia}, we can obtain an effective Hamiltonian containing interactions at all order between the visible nodes. In fact, it was showed that RBMs are universal approximator[4] of discrete distributions, that is, an arbitrary large RBM can approximate arbitrarily well any discrete distribution.

The learning procedure of the RBM is similar to the one of the BM, the difference lying in that the distribution is not in the exponential family anymore. Nevertheless, a gradient ascent can be achieved computing the derivative of the log-likelihood

$$\frac{\partial \mathscr{L}}{\partial w_{ia}} = \langle s_i \sum_{\tau_a=0,1} \tau_a p(\tau_a|\boldsymbol{s})\rangle_{\mathrm{d}} - \langle s_i \tau_a \rangle_{\mathscr{H}}$$
$$\frac{\partial \mathscr{L}}{\partial h_i} = \langle s_i \rangle_{\mathrm{d}} - \langle s_i \rangle_{\mathscr{H}}$$
$$\frac{\partial \mathscr{L}}{\partial \bar{h}_a} = \langle \sum_{\tau_a=0,1} \tau_a p(\tau_a|\boldsymbol{s})\rangle_{\mathrm{d}} - \langle \tau_a \rangle_{\mathscr{H}}$$

A notable difference is the presence of the conditioned average over the hidden node a in the term averaged over the dataset. In practice, it does not introduce more complication since the distribution conditioned over the value of the visible nodes factorizes

$$p(\tau_a = 1|\boldsymbol{s}) = \frac{1}{1+\exp\left(\sum_i w_{ia} s_i + \bar{h}_a\right)}$$

As a result, the training procedure for the RBM is similar to the one of the BM. At the opposite of the BM, the RBM likelihood is not convex anymore w.r.t. the parameters of the model and therefore the gradient ascent will in general not converge to the same set of parameters. An interesting outcomes of the bipartite architecture of the model is that, conditioned on one layer (visible or hidden), the marginal over the variables of the other layer factorizes. Using this properties we can implement an efficient sampling procedure: fixing the value of the variables of a given layer, the variables of the other layers can be drawn in parallel very quickly.

The RBM is a very complex object to analyze, both its equilibrium properties and the learning process. In the following, we will first show how a Gaussian RBM, despite being much more simple exhibits a non trivial learning behavior. Then we will show that, constraining the hidden layer to activate only

portamento di apprendimento non banale. Quindi mostreremo che, vincolando lo strato nascosto ad attivare un solo nodo alla volta, è possibile recuperare il cosiddetto *modello di miscele gaussiane* (MMG). Per questo modello, possiamo mostrare come l'apprendimento viene attivato dalle statistiche del set di dati, producendo una cascata di transizioni di fase. Infine, analizzeremo il comportamento dell'RBM, in campo medio, con variabili binarie ed in presenza di una matrice di peso strutturata. Si evincerà l'esistenza di una fase ferromagnetica, in qualche regione dello spazio dei parametri, il che potrà contribuire a spiegare come gli stati di equilibrio siano collegati al set di dati.

RBM gaussiana

In un primo tentativo di comprendere la dinamica di apprendimento del modello, possiamo guardare alla RBM gaussiana come un'estrema semplificazione dell'RBM. La RBM gaussiana (GRBM) consiste nell'usare una distribuzione gaussiana sia per le variabili visibili che per quelle nascoste, invece della distribuzione binaria $\{0,1\}$. La distribuzione quindi si legge

$$p_{GRBM}(\boldsymbol{s},\boldsymbol{\tau}) \propto p(\boldsymbol{s},\boldsymbol{\tau}) \prod_i e^{-\frac{s_i^2}{2\sigma_v^2}} \prod_a e^{-\frac{\tau_a^2}{2\sigma_h^2}} \qquad (5)$$

dove abbiamo definito le varianze intrinseche dei nodi visibili e nascosti come σ_v e σ_h. Dopo aver aggiunto le variabili nascoste, otteniamo una distribuzione gaussiana multi-variata sulle variabili visibili. Questo modello è interessante, non tanto per le sue proprietà di equilibrio quanto perché presenta dinamiche di apprendimento non banali che possono essere scritte esattamente [5, 6].

Per prima cosa, osserviamo che, usando la decomposizione ai valori singolari (i.e. Singular Value Decomposition, SVD) della matrice: $w_{ia} = \sum_\alpha u_i^\alpha w_\alpha v_a^\alpha$, definendo quindi l'insieme sinistro e destro di autovettori $\boldsymbol{u}^\alpha$ e $\boldsymbol{v}^\alpha$ di $\boldsymbol{w}$ insieme ai suoi autovalori w_α, possiamo diagonalizzare l'argomento dell'esponenziale di eq. (5). A tal fine, apportiamo la seguente modifica delle variabili

$$\hat{s}_\alpha = \sum_i s_i u_i^\alpha$$
$$\hat{\tau}_\alpha = \sum_a \tau_a v_a^\alpha$$

one hidden node at a time, we recover the so-called Gaussian mixtures model (GMM). For this model, we can show how the learning is triggered by the statistics of the dataset, yielding a cascade of phase transitions. Finally, we will analyze the mean-field behavior of the RBM with binary variables in presence of a structured weight matrix. The existence of a ferromagnetic phase is found in some region of the parameters space, possibly explaining how the equilibrium states are linked to the dataset.

Gaussian RBM

In a first attempt to understand the learning dynamics of the model, we can study the Gaussian RBM as an extreme simplification of the RBM. The Gaussian RBM consists in using a Gaussian distribution for both the visible and hidden variables, instead of the binary distribution $\{0,1\}$. The distribution hence reads

$$p_{GRBM}(\boldsymbol{s},\boldsymbol{\tau}) \propto p(\boldsymbol{s},\boldsymbol{\tau}) \prod_i e^{-\frac{s_i^2}{2\sigma_v^2}} \prod_a e^{-\frac{\tau_a^2}{2\sigma_h^2}} \qquad (5)$$

where we defined the intrinsic variance of the visible and hidden nodes as σ_v and σ_h. After summing the hidden variables, we get a multi-variate Gaussian distribution over the visible variables. Yet, this model is interesting, not because of its equilibrium properties but because it presents a non-trivial learning dynamics that can be written exactly[5, 6].

First, let's remark that, using the Singular Value Decomposition (SVD) of the matrix: $w_{ia} = \sum_\alpha u_i^\alpha w_\alpha v_a^\alpha$, hence defining the left and right set of eigenvectors $\boldsymbol{u}^\alpha$ and $\boldsymbol{v}^\alpha$ of $\boldsymbol{w}$ together with its eigenvalues w_α, we can diagonalize the argument of the exponential of eq. (5). To do so, we make the following change of variables

$$\hat{s}_\alpha = \sum_i s_i u_i^\alpha$$
$$\hat{\tau}_\alpha = \sum_a \tau_a v_a^\alpha$$

ed otteniamo la seguente distribuzione

$$p(\hat{s}) \propto \prod_{\alpha} \exp\left(-\frac{\hat{s}_\alpha^2}{2}\frac{1-\sigma_v^2\sigma_h^2 w_\alpha^2}{\sigma_v^2}\right) \qquad (6)$$

Osserviamo innanzitutto che, per una data matrice dei pesi, la distribuzione dell'eq. (6) descrive un insieme di variabili casuali gaussiane le cui varianze si dispongono lungo le direzioni principali di $\boldsymbol{w}$ date da $\sigma_\alpha^2 = \sigma_v^2/(1-\sigma_v^2\sigma_h^2 w_\alpha^2)$. Concentriamoci sulla traiettoria di apprendimento: proiettando le equazioni del gradiente sui modi singolari di $\boldsymbol{w}$, otteniamo la seguente espressione

$$\begin{aligned}\left(\frac{\partial \mathscr{L}}{\partial \boldsymbol{w}}\right)_{\alpha\beta} &= \sum_{ia} u_i^\alpha \frac{\partial \mathscr{L}}{\partial w_{ia}} v_a^\alpha \\ &= \langle \hat{s}_\alpha \hat{\tau}_\beta \rangle_{\mathrm{d}} - \langle \hat{s}_\alpha \hat{\tau}_\beta \rangle_{\mathscr{H}}\end{aligned}$$

Considerando un tasso di apprendimento infinitesimale η, possiamo identificare –nel limite di tempo continuo– $\frac{dw_{ia}}{dt} \sim \frac{\partial \mathscr{L}}{\partial w_{ia}}$. Calcolando la derivata temporale di ogni elemento della SVD della matrice dei pesi (u_i^α, w_α e v_a^α), otteniamo un altro insieme di equazioni del gradiente, questa volta sui modi singolari w_α e per le rotazioni infinitesime della matrice $\boldsymbol{u}$ e $\boldsymbol{v}$

$$\begin{aligned}\frac{dw_\alpha}{dt} &= \sigma_h^2 w_\alpha \left(\langle \hat{s}_\alpha^2 \rangle_{\mathrm{d}} - \frac{\sigma_v^2}{1-\sigma_v^2\sigma_h^2 w_\alpha^2}\right) \\ \Omega^v_{\alpha\beta} &= -\sigma_h^2 \left(\frac{w_\beta - w_\alpha}{w_\alpha + w_\beta} - \frac{w_\beta + w_\alpha}{w_\alpha - w_\beta}\right) \langle s_\alpha s_\beta \rangle_{\mathrm{d}} \\ \Omega^h_{\alpha\beta} &= -\sigma_h^2 \left(\frac{w_\beta - w_\alpha}{w_\alpha + w_\beta} + \frac{w_\beta + w_\alpha}{w_\alpha - w_\beta}\right) \langle s_\alpha s_\beta \rangle_{\mathrm{d}}\end{aligned}$$

dove $\Omega^{v,h}_{\alpha\beta}$ sono i generatori di rotazioni infinitesime per i vettori $\boldsymbol{u}^\alpha$, risp. $\boldsymbol{v}^\alpha$. È facile dedurre il comportamento a lungo termine di queste equazioni. I modi w_α convergeranno verso la seguente soluzione

$$w_\alpha^2 = \begin{cases} \frac{\langle \hat{s}_\alpha^2 \rangle_{\mathrm{d}} - \sigma_v^2}{\sigma_v^2 \sigma_h^2 \langle \hat{s}_\alpha^2 \rangle_{\mathrm{d}}} & \text{if } \langle \hat{s}_\alpha^2 \rangle_{\mathrm{d}} > \sigma_v^2 \\ 0 & \text{if } \langle \hat{s}_\alpha^2 \rangle_{\mathrm{d}} < \sigma_v^2 \end{cases} \qquad (7)$$

mostrando che se la loro varianza lungo la direzione principale α è inferiore alla varianza intrinseca dei nodi visibili, questi nodi verranno filtrati via. In caso contrario, il modo viene potenziato fino al valore dato nell'eq. (7): questo valore garantisce che la varianza nella direzione del modo α corrisponda alla varianza del set di dati nella stessa direzione. Il gradiente sulle rotazioni di $\boldsymbol{u}$ e $\boldsymbol{v}$ può essere annullato regolando le direzioni dei vettori $\boldsymbol{u}$ in modo tale che si diagonalizzi la matrice di covarianza del set di dati, ottenendo

and we obtain the following distribution

$$p(\hat{s}) \propto \prod_{\alpha} \exp\left(-\frac{\hat{s}_\alpha^2}{2}\frac{1-\sigma_v^2\sigma_h^2 w_\alpha^2}{\sigma_v^2}\right) \qquad (6)$$

Let us first observe that, for a given weight matrix, the distribution of eq. (6) describes an ensemble of Gaussian random variables of variance along the principal directions of $\boldsymbol{w}$ given by $\sigma_\alpha^2 = \sigma_v^2/(1-\sigma_v^2\sigma_h^2 w_\alpha^2)$. Let us focus on the learning trajectory. Projecting the equations of the gradient over the singular modes of $\boldsymbol{w}$, we obtain the following expression

$$\begin{aligned}\left(\frac{\partial \mathscr{L}}{\partial \boldsymbol{w}}\right)_{\alpha\beta} &= \sum_{ia} u_i^\alpha \frac{\partial \mathscr{L}}{\partial w_{ia}} v_a^\alpha \\ &= \langle \hat{s}_\alpha \hat{\tau}_\beta \rangle_{\mathrm{d}} - \langle \hat{s}_\alpha \hat{\tau}_\beta \rangle_{\mathscr{H}}\end{aligned}$$

Considering an infinitesimal learning rate η, we can identify in the continuous time limit that $\frac{dw_{ia}}{dt} \sim \frac{\partial \mathscr{L}}{\partial w_{ia}}$. Computing the time derivative of each element of the SVD of the weight matrix (u_i^α, w_α and v_a^α), we obtain another set of gradient equations, this time over the singular modes w_α and for the infinitesimal rotations of the matrix $\boldsymbol{u}$ and $\boldsymbol{v}$

$$\begin{aligned}\frac{dw_\alpha}{dt} &= \sigma_h^2 w_\alpha \left(\langle \hat{s}_\alpha^2 \rangle_{\mathrm{d}} - \frac{\sigma_v^2}{1-\sigma_v^2\sigma_h^2 w_\alpha^2}\right) \\ \Omega^v_{\alpha\beta} &= -\sigma_h^2 \left(\frac{w_\beta - w_\alpha}{w_\alpha + w_\beta} - \frac{w_\beta + w_\alpha}{w_\alpha - w_\beta}\right) \langle s_\alpha s_\beta \rangle_{\mathrm{d}} \\ \Omega^h_{\alpha\beta} &= -\sigma_h^2 \left(\frac{w_\beta - w_\alpha}{w_\alpha + w_\beta} + \frac{w_\beta + w_\alpha}{w_\alpha - w_\beta}\right) \langle s_\alpha s_\beta \rangle_{\mathrm{d}}\end{aligned}$$

where $\Omega^{v,h}_{\alpha\beta}$ are the generator of infinitesimal rotations for the vectors $\boldsymbol{u}^\alpha$, resp. $\boldsymbol{v}^\alpha$. It is easy to deduce the long time behavior of these equations. The modes w_α will converge toward the following solution

$$w_\alpha^2 = \begin{cases} \frac{\langle \hat{s}_\alpha^2 \rangle_{\mathrm{d}} - \sigma_v^2}{\sigma_v^2 \sigma_h^2 \langle \hat{s}_\alpha^2 \rangle_{\mathrm{d}}} & \text{if } \langle \hat{s}_\alpha^2 \rangle_{\mathrm{d}} > \sigma_v^2 \\ 0 & \text{if } \langle \hat{s}_\alpha^2 \rangle_{\mathrm{d}} < \sigma_v^2 \end{cases} \qquad (7)$$

showing that if the variance along the principal direction α is lower than the intrinsic variance of the visible nodes, the mode is filtered out. Otherwise, the mode is enhanced up to the value given in eq. (7). This value ensure that the variance in the direction of the mode α matches the variance of the dataset in the same direction. The gradient over the rotations of $\boldsymbol{u}$ and $\boldsymbol{v}$ can be canceled by adjusting the directions of the vectors $\boldsymbol{u}$ such that it diagonalizes the covariance matrix of the dataset, giving $\langle s_\alpha s_\beta \rangle_{\mathrm{d}} = 0$ if $\alpha \neq \beta$.

La Decomposizione ai Valori Singolari (SVD)

La SVD è una fattorizzazione di una matrice, che generalizza la diagonalizzazione a matrici rettangolari. Data una matrice $\boldsymbol{M}$ di dimensione $n \times m$, è **sempre** possibile scomporla su un insieme di autovettori sinistro (destro): $M_{ij} = \sum_\alpha u_i^\alpha s_\alpha v_j^\alpha$, dove il numero di tali vettori è dato da $\min(n,m)$. Nella notazione matriciale scriviamo $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}$, dove $\boldsymbol{\Sigma}$ ha componenti diverse da zero solo sui suoi elementi diagonali. L'insieme dei vettori $\boldsymbol{u}^\alpha$ (risp. $\boldsymbol{v}^\alpha$) per una base ortogonale, $\sum_i u_i^\alpha u_i^\beta = \delta_{\alpha\beta}$ (risp. $\sum_j v_j^\alpha v_j^\beta = \delta_{\alpha\beta}$), ed i singoli valori s_α sono tutti positivi o nulli. La matrice $\boldsymbol{M}$ ha le seguenti proprietà: $\boldsymbol{u}^\alpha \boldsymbol{M} = s_\alpha \boldsymbol{M}\boldsymbol{v}^\alpha$ e $\boldsymbol{M}\boldsymbol{u}^\alpha = s_\alpha \boldsymbol{v}^\alpha \boldsymbol{M}$. Quando $\boldsymbol{M}$ corrisponde a un dataset centrato (le colonne rappresentano punti in uno spazio m-dimensionale), possiamo vedere che la SVD è collegata all'analisi delle componenti principali (PCA): $(1/m)\boldsymbol{M}\boldsymbol{M}^T = \boldsymbol{U}\Sigma^2\boldsymbol{U}^T$, dove la matrice $\boldsymbol{U}$ corrisponde alle direzioni principali e la matrice diagonale $\boldsymbol{\Sigma}^2$ ai valori principali.

The Singular Value Decomposition (SVD)

The SVD is a factorization of a matrix, generalizing the eigendecomposition to rectangular matrix. For a given matrix $\boldsymbol{M}$ of size $n \times m$, it is **always** possible to decompose it on a set of left(right) eigenvector : $M_{ij} = \sum_\alpha u_i^\alpha s_\alpha v_j^\alpha$, where the number of such vector is given by $\min(n,m)$. In matrix notation it is written $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}$, where $\boldsymbol{\Sigma}$ has non-zero components only on its diagonal elements. The set of vectors $\boldsymbol{u}^\alpha$ (resp. $\boldsymbol{v}^\alpha$) for an orthogonal basis: $\sum_i u_i^\alpha u_i^\beta = \delta_{\alpha\beta}$ (resp. $\sum_j v_j^\alpha v_j^\beta = \delta_{\alpha\beta}$), and the singular values s_α are all positive or zero. The matrix $\boldsymbol{M}$ has the following properties : $\boldsymbol{u}^\alpha \boldsymbol{M} = s_\alpha \boldsymbol{M}\boldsymbol{v}^\alpha$ and $\boldsymbol{M}\boldsymbol{u}^\alpha = s_\alpha \boldsymbol{v}^\alpha \boldsymbol{M}$. When $\boldsymbol{M}$ correspond to a centered dataset (the columns represent points in a m-dimensional space), we can see that the SVD is linked to the principal components analysis (PCA): $1/m\boldsymbol{M}\boldsymbol{M}^T = \boldsymbol{U}\Sigma^2\boldsymbol{U}^T$, where the matrix $\boldsymbol{U}$ correspond to the principal directions and the diagonal matrix $\boldsymbol{\Sigma}^2$ to the principal values.

$\langle s_\alpha s_\beta \rangle_{\mathrm{d}} = 0$ se $\alpha \neq \beta$.

Pertanto, per la RBM gaussiana:

- la dinamica di apprendimento ruotano la matrice $\boldsymbol{u}$ fino a trovare le direzioni principali del set di dati;
- questa evidenzierà anche i modi per i quali la varianza lungo la corrispondente direzione principale è maggiore della varianza intrinseca delle variabili visibili;
- il valore assunto da un modo α è tale che la varianza della distribuzione appresa verso quella direzione corrisponda alla varianza del set di dati nella stessa direzione.

Poiché ci aspettiamo che il regime lineare qui descritto avvenga all'inizio del processo di apprendimento anche per macchine più complesse, il comportamento ottenuto dovrebbe darci qualche suggerimento sulle dinamiche di apprendimento per la RBM non lineare.

Softmax RBM

Therefore, for the Gaussian RBM:

- the learning dynamics will rotate the matrix $\boldsymbol{u}$ until it finds the principal directions of the dataset;
- it will also express the modes for which the variance along the corresponding principal direction is higher than the intrinsic variance of the visible variables;
- the value taken by a mode α is such that the variance of the learned distribution toward that direction matches the variance of the dataset in the same direction.

Since we expect the linear regime described here to take place at the beginning of learning process for more complex machine, the obtained behavior should give us some hint on the learning dynamics for non-linear RBM.

Softmax RBM

Prima di analizzare la RBM con variabili $\{0,1\}$ discrete, focalizziamo la nostra attenzione su una RBM di complessità intermedia dove i nodi visibili seguono una prior gaussiana mentre i nodi nascosti rimangono discreti in $\{0,1\}$, ma con il vincolo di avere un solo nodo attivabile. Chiamiamo softmax-RBM questa RBM poiché i nodi nascosti seguono la distribuzione softmax. Per questo modello, per calcolare la distribuzione marginale sui nodi visibili, diamo prima la distribuzione di probabilità condizionata sui nodi nascosti

$$p(\tau_a = 1, \tau_{b\neq a} = 0|\boldsymbol{s}) = \frac{\exp\left(\sum_i w_{ia}s_i + \bar{h}_a\right)}{\sum_b \exp\left(\sum_i w_{ib}s_i + \bar{h}_b\right)}$$

Quando calcoliamo la distribuzione marginale sul nodo visibile riconosciamo la distribuzione del modello di miscele gaussiane (GMM) [7, 8]

$$p(\boldsymbol{s}) = \frac{1}{Z}\sum_a \rho_a \exp\left(\sum_i -\frac{1}{2\sigma_v^2}(s_i - w_{ia})^2\right)$$

dove abbiamo assorbito il campo magnetico h_i nella definizione dei pesi e lo abbiamo riscalato di σ_v^2 per disaccoppiare la varianza intrinseca dal centro dei modi gaussiani $w'_{ia} = \sigma_v^2(w_{ia} + h_i)$. Vediamo che la matrice dei pesi $\boldsymbol{w}'$ rappresenta il centro delle componenti gaussiane e la densità corrispondente è data da

$$\rho_a = \frac{\exp\left(\bar{h}_a + \sum_i w'^2_{ia}\right)}{\sum_b \exp\left(\bar{h}_b + \sum_i w'^2_{ia}\right)}$$

In questa formulazione, l'interpretazione dei nodi nascosti è chiara: per una data configurazione visibile $\boldsymbol{s}$, il nodo nascosto corrispondente alla distribuzione gaussiana "più vicina " avrà la più alta probabilità di essere attivato, mentre, per un nodo nascosto a attivato, la corrispondente configurazione visibile è data da una gaussiana centrata su w'_{ia} con varianza σ_v^2.

Come nella RBM gaussiana, gli aspetti interessanti di questo modello non sono le proprietà di equilibrio che sono banali, ma, piuttosto, le dinamiche di apprendimento. A seguire mostriamo che, variando la varianza intrinseca del sistema (che può essere vista come la temperatura), il sistema subisce una transizione di fase da una fase "paramagnetica" –dove i centri w'_{ia} delle gaussiane vengono adattati al centro di massa del set di dati– verso un'altra fase –in cui questi centri diffondono in punti diversi del set di dati. Per

Before analyzing the RBM with discrete $\{0,1\}$ variables, we focus our attention on an RBM of intermediate complexity where the visible nodes follow a Gaussian prior and the hidden nodes will be discrete in $\{0,1\}$ with the constraint of having only one node that can be activated. This RBM will be called the softmax-RBM since the hidden nodes follow the softmax distribution. For this model, in order to compute the marginal over the visible nodes, let us first give the conditioned probability distribution over the hidden nodes

$$p(\tau_a = 1, \tau_{b\neq a} = 0|\boldsymbol{s}) = \frac{\exp\left(\sum_i w_{ia}s_i + \bar{h}_a\right)}{\sum_b \exp\left(\sum_i w_{ib}s_i + \bar{h}_b\right)}$$

The marginal over the visible node is then computed and we recognize the distribution of the Gaussian mixtures model[7, 8]

$$p(\boldsymbol{s}) = \frac{1}{Z}\sum_a \rho_a \exp\left(\sum_i -\frac{1}{2\sigma_v^2}(s_i - w_{ia})^2\right)$$

where we absorbed the magnetic field h_i in the definition of the weights and re-scaled it by σ_v^2 in order to decouple the intrinsic variance from the center of the Gaussian modes $w'_{ia} = \sigma_v^2(w_{ia} + h_i)$. We see that the weights matrix $\boldsymbol{w}'$ represent the center of the Gaussian components and the corresponding density is given by

$$\rho_a = \frac{\exp\left(\bar{h}_a + \sum_i w'^2_{ia}\right)}{\sum_b \exp\left(\bar{h}_b + \sum_i w'^2_{ia}\right)}$$

In this formulation, the interpretation of the hidden nodes is clear. For a given visible configuration $\boldsymbol{s}$, the hidden node corresponding to the "closest" Gaussian distribution will have the highest probability to be turned on. In the other way around, for a given activated hidden node a, the corresponding visible configuration is given by a Gaussian centered on w'_{ia} with variance σ_v^2.

As in the Gaussian RBM, the interesting aspects of this model is not the equilibrium properties which are trivial, but rather the learning dynamics. We will show that, by varying the intrinsic variance of the system (which can be seen as the temperature), the system undergoes a phase transition from a "paramagnetic" phase, where the centers w'_{ia} of the Gaussian are adjusted to the center of mass of the dataset to another phase where the centers of the Gaussian spread in different places of the dataset. First, let's write the

prima cosa, scriviamo le equazioni di apprendimento: derivando la verosimiglianza del sistema, otteniamo la seguente espressione per i due termini del gradiente

$$\langle s_i \tau_a \rangle_{\mathrm{d}} = \frac{1}{M} \sum_d \left(s_i^{(d)} - w'_{ia} \right) p(\tau_a | \boldsymbol{s}^{(d)}) \qquad (8)$$

$$\langle s_i \tau_a \rangle_{\mathscr{H}} = \frac{1}{M} \sum_d w'_{ia} \left[p(\tau_a | \boldsymbol{s}^{(d)}) - \rho_a \right] \qquad (9)$$

Per coloro che hanno familiarità con il GMM, osserviamo che l'eq. (8) può essere trasformata nello schema di iterazione EM dell'equazione di apprendimento per il GMM imponendo che il lato sinistro sia nullo e che la distribuzione condizionata $p(\tau_a|\boldsymbol{s})$ non dipenda da w'_{ia}: l'eq. (9) aggiusterà la densità ρ_a di ciascuna gaussiana secondo una misura di densità locale. Analizziamo il comportamento del processo di apprendimento ad alta σ_v: innanzitutto, assumiamo che il set di dati sia stato centrato, avendo $\sum_d s_i^{(d)} = 0$, $\forall i$. Quindi prendiamo, come condizione iniziale, un valore di σ_v che sia grande rispetto alla varianza del set di dati in qualsiasi direzione. Nel limite di varianza infinita, è chiaro che una soluzione banale delle equazioni di apprendimento è data da $w'_{ia} = 0$, $\bar{h}_a = 0$, e quindi $\rho_a = 1/N_h$ e $p(\tau_a|\boldsymbol{s}) = 1/N_h$ per tutti i nodi nascosti. In altre parole, ogni configurazione visibile ha la stessa probabilità di essere assegnata a qualsiasi gaussiana e tutte le gaussiane hanno la stessa densità *a priori* ρ_a: il modello ha appreso solo il centro di massa del set di dati. Per studiare cosa succede quando la varianza è diminuita, possiamo studiare la stabilità lineare della soluzione paramagnetica, aggiungendo una piccola perturbazione ai centri: $w_{ia} = \varepsilon_{ia}$ ed indagando il comportamento della dinamica del gradiente

$$w_{ia}^{\prime(t+1)} = w_{ia}^{\prime(t)} - \eta \frac{1}{M} \sum_d \left(s_i^{(d)} - w_{ia}^{\prime(t)} \right) p(\tau_a | \boldsymbol{s}^{(d)})$$

Linearizzando le equationi all'ordine ε otteniamo

$$\varepsilon_{ia}^{(t+1)} = (1-\eta)\varepsilon_{ia}^{(t)} + \frac{\eta}{\sigma_v^2} \sum_j c_{ij} \left(\varepsilon_{ja}^{(t)} - \frac{1}{N_h} \sum_b \varepsilon_{jb}^{(t)} \right).$$

Si vede che, non appena l'autovalore massimo della matrice di covarianza Γ_C è maggiore di σ_v^2, la soluzione paramagnetica diventa instabile anche per piccole fluttuazioni. Questo innesca l'apprendimento in maniera tale che la posizione dei centri si allontanerà dal centro di massa seguendo la prima direzione principale della matrice di covarianza: poiché il gradiente

learning equations. By deriving the likelihood of the system, we obtain the following expression for the two terms in the gradient

$$\langle s_i \tau_a \rangle_{\mathrm{d}} = \frac{1}{M} \sum_d \left(s_i^{(d)} - w'_{ia} \right) p(\tau_a | \boldsymbol{s}^{(d)}) \qquad (8)$$

$$\langle s_i \tau_a \rangle_{\mathscr{H}} = \frac{1}{M} \sum_d w'_{ia} \left[p(\tau_a | \boldsymbol{s}^{(d)}) - \rho_a \right] \qquad (9)$$

For those familiar with the GMM, we remark that eq. (8) can be turned into the Expectation-Maximization iteration scheme of the learning equation for the GMM by imposing the l.h.s. is zero, and that the conditioned distribution $p(\tau_a|\boldsymbol{s})$ does not depend on w'_{ia}. The eq. (9) will adjust the density ρ_a of each Gaussian according to a local density measure. Let's investigate the behavior of the learning process at high σ_v. First, we consider that the dataset has been centered, having $\sum_d s_i^{(d)} = 0$, $\forall i$. Then we take, as initial condition, a value of σ_v which is large in comparison to the variance of the dataset in any direction. In the infinite variance limit, it is clear that a trivial solution of the learning equations is given by $w'_{ia} = 0$, $\bar{h}_a = 0$, and thus $\rho_a = 1/N_h$, and $p(\tau_a|\boldsymbol{s}) = 1/N_h$ for all hidden nodes. In other words, a visible configuration has the same probability to be assigned to any of the Gaussian. And all the Gaussian have the same *a priori* density ρ_a. The model learned only the center of mass of the dataset. To study what happened when the variance is decreased, we can study the linear stability of the paramagnetic solution, by adding a small perturbation to the centers : $w_{ia} = \varepsilon_{ia}$ and investigating the behavior of the gradient dynamics

$$w_{ia}^{\prime(t+1)} = w_{ia}^{\prime(t)} - \eta \frac{1}{M} \sum_d \left(s_i^{(d)} - w_{ia}^{\prime(t)} \right) p(\tau_a | \boldsymbol{s}^{(d)})$$

Linearizing the equations at the order ε we obtain

$$\varepsilon_{ia}^{(t+1)} = (1-\eta)\varepsilon_{ia}^{(t)} + \frac{\eta}{\sigma_v^2} \sum_j c_{ij} \left(\varepsilon_{ja}^{(t)} - \frac{1}{N_h} \sum_b \varepsilon_{jb}^{(t)} \right).$$

where c_{ij} is empirical covariance matrix of the dataset. We see that, as soon as the maximum eigenvalue of the covariance matrix Γ_C is higher than σ_v^2, the paramagnetic solution becomes unstable to small fluctuations. It triggers the learning where the position of the centers will move away from the center of mass following the first principal direction of the

rappresenta la prima derivata della verosimiglianza (o equivalentemente dell'energia libera) del sistema, vediamo che il sistema subisce una transizione di fase, trovando altri minimi stabili.

Sorprendentemente, la transizione di fase in gioco qui è molto simile al comportamento della RBM gaussiana: in entrambi i modelli, l'apprendimento è innescato dalle proprietà della matrice di covarianza. Per la RBM gaussiana sono espressi tutti i modi aventi varianze maggiori delle varianze intrinseche. Nel softmax-RBM, il modo principale più forte attiva una divisione dei centri, diffondendoli lungo la direzione principale. È facile convincersi che questo fenomeno appare in modo gerarchico, poiché la varianza intrinseca sta diminuendo sempre di più. Come notevole differenza tra i due modelli, il softmax-RBM appreso può essere multimodale alla fine dell'apprendimento.

Binary RBM

Passare dal modello semplice a quello più complesso ci aiuta ad avere un'intuizione su quale potrebbe essere il comportamento di quello più complesso. Nel caso della RBM, dobbiamo naturalmente aspettarci che, partendo da un regime di "alta temperatura" con un minimo globale unico di energia libera, il meccanismo di apprendimento spingerà il sistema verso un'altra fase, scindendosi in una descrizione multimodale del set di dati. Utilizzando il comportamento di apprendimento della RBM gaussiana come guida, dobbiamo aspettarci che inizialmente che la RBM apprenda la SVD del set di dati. Successivamente, la non linearità sarà non trascurabile e la dinamica risultante sarà molto più difficile da analizzare.

Per questo modello, ci concentriamo prima sul comportamento di equilibrio della RBM. La difficoltà nell'analizzare questo regime è che gli strumenti analitici tradizionali (come il metodo delle repliche [9]) si basano sull'indipendenza dei singoli accoppiamenti (a dire, degli ingressi della matrice dei pesi). Nella RBM è chiaro che il processo di apprendimento introduce una forte correlazione tra gli elementi della matrice dei pesi. Tuttavia, un primo approccio, utilizzando una matrice diluita di elementi indipendenti, mostra che una tale RBM può mostrare una fase interessante in cui le caratteristiche apprese sono opportunamente *composte* per richiamare uno schema memorizzato [10, 11]. Questo approccio è tuttavia molto complesso e ne preferiremo qui un altro [6]

covariance matrix. Since the gradient represent the first derivative of the likelihood (or equivalently of the free energy) of the system, we see that the system will undergo a phase transition finding other stable minima.

Remarkably, the phase transition at stake here is very similar to the behavior of the Gaussian RBM: in both models, the learning is triggered by the properties of the covariance matrix. For the Gaussian RBM all the modes having variances higher than the intrinsic variances are expressed. In the softmax-RBM, the strongest principal mode will trigger a split of the centers, scattering them along the principal direction. It is easy to be convinced that this phenomena will appear in a hierarchical manner, as the intrinsic variance is decreasing more and more. As notable difference between the two models, the learned softmax-RBM can be multi-modal at the end of the learning.

Binary RBM

Going from simple model to more complex ones help us to have an intuition about what could be the behavior of the more complex one. In the case of RBM, we shall naturally expect that, starting in a "high temperature" regime with a unique global minimum of the free energy, the learning mechanism will push the system toward another phase, splitting into a multimodal description of the dataset. Using the learning behavior of the Gaussian RBM, we should expect at first that the RBM will learn SVD of the dataset. Later on, non-linearity will be non-negligible and the resulting dynamics will be much harder to analyze.

For this model, we focus first on the equilibrium behavior of the RBM. The difficulty to analyze this regime is that, traditional analytical tools (such as the replica method [9]) rely on the independence of the elements of the coupling or weight matrix. In the RBM, it is clear that the learning process introduces strong correlation among the elements of the weight matrix. Still, a first approach, using a diluted matrix of independent elements, shows that such an RBM can show an interesting phase where the learned features are composed to recall a memorized pattern [10, 11]. This approach is however very complex and we shall prefer here another one [6] based on a different construction of the weight matrix highlighting the importance of the SVD of the matrix. This construction relies on the hypothesis that the weight matrix contains a structured part of rank $K = \mathcal{O}(1)$ in

basato su una diversa costruzione della matrice dei pesi. Questa costruzione si basa sull'ipotesi che la matrice dei pesi contenga una parte strutturata di rango $K = \mathcal{O}(1)$ oltre ad una matrice casuale corrispondente al rumore:

$$w_{ia} = \sum_{\alpha=1}^{K} u_i^\alpha w_\alpha v_a^\alpha + r_{ia}$$

dove $K \ll N_v$, assumendo quindi una scomposizione di basso rango della matrice dei pesi più un rumore gaussiano casuale r_{ia} di varianza σ. Diamo qui uno schizzo dei diversi passaggi salienti per calcolare l'energia libera del sistema utilizzando il metodo delle repliche. Innanzitutto, la nostra ipotesi è che il modo w_α della matrice dei pesi rappresenti alcune proprietà intrinseche apprese da un insieme di dati, mentre i vettori $\boldsymbol{u}^\alpha$, $\boldsymbol{v}^\alpha$ e $\boldsymbol{r}$ corrispondano al disordine congelato (i.e. *quenched*). Sotto questo assunto, dobbiamo calcolare l'energia libera quenched $F = \mathbb{E}(\log Z)$, dove $\mathbb{E}(.)$ rappresenta la media sulla variabile quenched. A tale scopo si presta il metodo delle repliche, basato sulla seguente identità

$$\log Z = \lim_{n\to 0} \frac{Z^n - 1}{n}$$

con la speranza che il calcolo della media quenched di Z^n, e successivamente il limite $n \to 0$, portino al risultato corretto. Per calcolare questa quantità, introduciamo le variabili replicate

$$Z^n = \sum_{\{\boldsymbol{s}^1,\boldsymbol{\tau}^1\},\{\boldsymbol{s}^2,\boldsymbol{\tau}^2\},\ldots,\{\boldsymbol{s}^n,\boldsymbol{\tau}^n\}} \exp\left(\sum_{i,a,p} s_i^p w_{ia} \tau_a^p\right).$$

Per eseguire l'integrazione sulla matrice gaussiana casuale $\boldsymbol{r}$ fattorizziamo tutto il termine proporzionale a w_{ia}: l'integrale introduce un accoppiamento efficace tra le repliche

$$\int \mathcal{D}w_{ia} e^{w_{ia}\sum_p s_i^p \tau_a^p} = \exp\left[\frac{\sigma^2}{2L} \sum_{i,a,p\neq q} s_i^p s_i^q \tau_a^p \tau_a^q\right].$$

Questa interazione tra nodi visibili e nascosti può essere disaccoppiata usando la trasformazione di Hubbard-Stratonovitch (HS) [12, 13]

$$\exp(xy) = \int d\bar{x}d\bar{y} e^{-\bar{x}\bar{y}+\bar{x}y+\bar{y}x}$$

dove i nuovi parametri $\bar{x}$ (risp. $\bar{y}$) sono il coniugato di x (risp. y). A seguire introduciamo i parametri d'ordine del vetro di spin Q_{pq} e $\bar{Q}_{pq}$, coniugati rispet-

addition to a random matrix corresponding to noise:

$$w_{ia} = \sum_{\alpha=1}^{K} u_i^\alpha w_\alpha v_a^\alpha + r_{ia}$$

where $K \ll N_v$, assuming a low-rank decomposition of the weight matrix plus some random gaussian noise r_{ia} of variance σ. We give here a sketch of the different steps in order to compute the free energy of the system using the replica approach. First, our hypothesis is that the mode w_α of the weight matrix represents some learned intrinsic properties of a dataset, while the vectors $\boldsymbol{u}^\alpha$, $\boldsymbol{v}^\alpha$ and $\boldsymbol{r}$ correspond to quenched disorder. Under this assumption, we need to compute the quenched free energy $F = \mathbb{E}(\log Z)$, where $\mathbb{E}(.)$ represent the average over the quenched variable. To do that, we will use the replica method based upon the following identity

$$\log Z = \lim_{n\to 0} \frac{Z^n - 1}{n}$$

with the hope that computing the quenched average of Z^n, and taking the limit $n \to 0$ afterward, leads to the correct result. To compute this quantity, we introduce the replicated variables s_i^p and τ_a^p, where p indicates replica's indices, leading to

$$Z^n = \sum_{\{\boldsymbol{s}^1,\boldsymbol{\tau}^1\},\{\boldsymbol{s}^2,\boldsymbol{\tau}^2\},\ldots,\{\boldsymbol{s}^n,\boldsymbol{\tau}^n\}} \exp\left(\sum_{i,a,p} s_i^p w_{ia} \tau_a^p\right)$$

In order to perform the integral over the random Gaussian matrix $\boldsymbol{r}$ we factorize all the term proportional to w_{ia}. The integral introduces an effective coupling between the replicas

$$\int \mathcal{D}w_{ia} e^{w_{ia}\sum_p s_i^p \tau_a^p} = \exp\left[\frac{\sigma^2}{2L} \sum_{i,a,p\neq q} s_i^p s_i^q \tau_a^p \tau_a^q\right]$$

This interaction between the visible and the hidden nodes can be decoupled by using the Hubbard-Stratonovitch [12, 13] (HS) transformation

$$\exp(xy) = \int d\bar{x}d\bar{y} e^{-\bar{x}\bar{y}+\bar{x}y+\bar{y}x}$$

where the new parameters $\bar{x}$ (resp. $\bar{y}$) are the conjugate of x (resp. y). In our case, we introduce the spin glass order parameters Q_{pq} and $\bar{Q}_{pq}$, conjugate

tivamente di $\sum_a \tau_a^p \tau_a^q$ e $\sum_i s_i^p s_i^q$. Continuiamo ad applicare nuovamente la trasformazione HS ai seguenti termini

$$\sum_{i,a,\alpha} s_i^p u_i^\alpha w_\alpha v_a^\alpha \tau_a^p = \sum_\alpha w_\alpha (\sum_i s_i^p u_i^\alpha)(\sum_a \tau_a^p v_a^\alpha)$$

introducendo i parametri d'ordine m_α^p e $\bar{m}_\alpha^p$ coniugato con $\tau_\alpha = (1/\sqrt{L}) \sum_a \tau_a^p v_a^\alpha$ e $s_\alpha = (1/\sqrt{L}) \sum_i s_i^p u_i^\alpha$. Infine, dobbiamo sommare sulle variabili $\{\boldsymbol{s}^p\}$ e $\{\boldsymbol{\tau}^p\}$ e calcolare la media sulle matrici $\boldsymbol{u}$ e $\boldsymbol{v}$. Per fare ciò, assumiamo che gli elementi delle matrici $\boldsymbol{u}$ e $\boldsymbol{v}$ siano distribuiti in modo identico ed indipendente, senza per ora specificare la loro distribuzione. Con questa semplificazione, possiamo fattorizzare i termini dipendenti dagli indici visibili o nascosti i, a. Infine, assumiamo l'ipotesi di simmetria di replica: $Q_{pq} = q$, $\bar{Q}_{pq} = \bar{q}$, $m_\alpha^p = m_\alpha$ e $\bar{m}_\alpha^p = \bar{m}_\alpha$. Dopo aver preso il limite termodinamico $N_v, N_h \to \infty$, mantenendo costante il rapporto $\kappa = \sqrt{N_h/N_v}$ e, prendendo il limite $n \to 0$, otteniamo un'espressione esplicita per l'energia libera quenched

$$f[m,\bar{m},q,\bar{q}] = \sum_\alpha w_\alpha m_\alpha \bar{m}_\alpha - \frac{\sigma^2}{2} q\bar{q} + \frac{\sigma^2}{2}(q+\bar{q})$$
$$-\frac{1}{\sqrt{\kappa}} \mathbb{E}_{u,x}[\cosh\big(h(x,u)\big)] - \sqrt{\kappa}\, \mathbb{E}_{v,x}[\cosh\big(\bar{h}(x,v)\big)]$$

con $\kappa = \sqrt{N_h/N_v}$, essendo $\mathbb{E}$ l'operatore di media (x è una variabile stocastica Gaussiana centrata di varianza unitaria) e da cui è immediato ricavare le equazioni di punto sella

$$m_\alpha = \kappa^{\frac{1}{4}} \mathbb{E}_{v,x}\Big[v^\alpha \tanh\big(\bar{h}(x,v)\big)\Big],$$
$$q = \mathbb{E}_{v,x}\Big[\tanh^2\big(\bar{h}(x,v)\big)\Big]$$

e gli associati $\bar{m}_\alpha$ e $\bar{q}$ ottenuti mandando $\bar{h} \to h$. Le funzioni h e $\bar{h}$ sono date da

$$h(x,u) \stackrel{\text{def}}{=} \kappa^{\frac{1}{4}} \big(\sigma\sqrt{q}x + \sum_\gamma w_\gamma m_\gamma u^\gamma\big)$$

$$\bar{h}(x,v) \stackrel{\text{def}}{=} \kappa^{-\frac{1}{4}} \big(\sigma\sqrt{\bar{q}}x + \sum_\gamma w_\gamma \bar{m}_\gamma v^\gamma\big),$$

L'interpretazione dei parametri dell'ordine è chiara. m_α (risp. $\bar{m}_\alpha$) corrisponde alla magnetizzazione visibile (risp. nascosta) proiettata nella direzione del modo α. Pertanto un valore diverso da zero indica che la magnetizzazione di equilibrio è altamente correlata con un dato modo. q (risp. $\bar{q}$) sono i parametri d'ordine del vetro di spin, che indicano se il sistema

respectively of $\sum_a \tau_a^p \tau_a^q$ and $\sum_i s_i^p s_i^q$. We continue applying again the HS transformation to the following terms

$$\sum_{i,a,\alpha} s_i^p u_i^\alpha w_\alpha v_a^\alpha \tau_a^p = \sum_\alpha w_\alpha (\sum_i s_i^p u_i^\alpha)(\sum_a \tau_a^p v_a^\alpha)$$

introducing the order parameters m_α^p and $\bar{m}_\alpha^p$ conjugate of $\tau_\alpha = 1/\sqrt{L}\sum_a \tau_a^p v_a^\alpha$ and $s_\alpha = 1/\sqrt{L}\sum_i s_i^p u_i^\alpha$. Finally, we need to sum over the variables $\{\boldsymbol{s}^p\}$ and $\{\boldsymbol{\tau}^p\}$ and to average over the matrices $\boldsymbol{u}$ and $\boldsymbol{v}$. To do this, we assume that the elements of the matrices $\boldsymbol{u}$ and $\boldsymbol{v}$ are independent identically distributed (i.i.d.), without so far specifying their distribution. With this simplification, we can factorize the terms dependent on the visible or hidden indices i,a. Finally, we make the replica-symmetric hypothesis: $Q_{pq} = q$, $\bar{Q}_{pq} = \bar{q}$, $m_\alpha^p = m_\alpha$ and $\bar{m}_\alpha^p = \bar{m}_\alpha$. After taking the thermodynamics limit $N_v, N_h \to \infty$, keeping the ratio $\kappa = \sqrt{N_h/N_v}$ constant and, taking the limit $n \to 0$, we obtain the quenched free energy

$$f[m,\bar{m},q,\bar{q}] = \sum_\alpha w_\alpha m_\alpha \bar{m}_\alpha - \frac{\sigma^2}{2} q\bar{q} + \frac{\sigma^2}{2}(q+\bar{q})$$
$$-\frac{1}{\sqrt{\kappa}} \mathbb{E}_{u,x}[\cosh\big(h(x,u)\big)] - \sqrt{\kappa}\, \mathbb{E}_{v,x}[\cosh\big(\bar{h}(x,v)\big)]$$

with $\kappa = \sqrt{N_h/N_v}$, $\mathbb{E}$ being the average over the corresponding variables (x is a Gaussian centered stochastic variable of unit variance) and where the saddle point equations are given by

$$m_\alpha = \kappa^{\frac{1}{4}} \mathbb{E}_{v,x}\Big[v^\alpha \tanh\big(\bar{h}(x,v)\big)\Big],$$
$$q = \mathbb{E}_{v,x}\Big[\tanh^2\big(\bar{h}(x,v)\big)\Big]$$

and the associated $\bar{m}_\alpha$ and $\bar{q}$ obtained by changing $\bar{h} \to h$. The function h and $\bar{h}$ are given by

$$h(x,u) \stackrel{\text{def}}{=} \kappa^{\frac{1}{4}} \big(\sigma\sqrt{q}x + \sum_\gamma w_\gamma m_\gamma u^\gamma\big)$$

$$\bar{h}(x,v) \stackrel{\text{def}}{=} \kappa^{-\frac{1}{4}} \big(\sigma\sqrt{\bar{q}}x + \sum_\gamma w_\gamma \bar{m}_\gamma v^\gamma\big),$$

The interpretation of the order parameters is clear. The m_α (resp. $\bar{m}_\alpha$) corresponds to the visible (resp. hidden) magnetization projected in the direction of the mode α. Therefore a non-zero value indicates that the equilibrium magnetization is highly correlated with a given mode. The q (resp. $\bar{q}$) is the spin-glass order parameters, indicating whether or not the

possa rimanere intrappolato o meno in un insieme di configurazioni simili. Queste equazioni di campo medio sono soddisfatte nei minimi dell'energia libera del sistema. Possiamo ora concentrarci sulle diverse fasi del sistema. Seguendo l'analisi tradizionale della teoria dei vetri di spin, possiamo cercare la stabilità delle seguenti regioni

- $m_\alpha = \bar{m}_\alpha = q = \bar{q} = 0$ **la fase paramagnetica**. Si verifica in genere ad alta temperatura (per piccoli σ) ed accoppiamenti deboli (piccoli w_α).
- $m_\alpha, \bar{m}_\alpha, q, \bar{q} \neq 0$ **la fase ferromagnetica**. In questa fase, ci aspettiamo che il sistema condensi sui modi singolari: le configurazioni di equilibrio hanno una sovrapposizione macroscopica con uno o più di questi modi.
- $m_\alpha = \bar{m}_\alpha = 0$ e $q, \bar{q} \neq 0$ **fase di vetro di spin**. In questa fase, il sistema è bloccato in poche configurazioni a basso consumo energetico ma queste configurazioni non sono correlate ai modi singolari di $\boldsymbol{w}$

Per recuperare configurazioni correlate ai segnali w_α è necessario entrare nella fase ferromagnetica: intuitivamente possiamo supporre che quando la matrice del rumore è debole (cioè per piccoli valori di σ) ed anche il segnale è debole (piccolo w_α) la rete sia nella fase paramagnetica. Assumendo si voglia evitare la fase di vetro di spin ad esempio, la domanda da porsi è capire cosa possa far prevalere la fase ferromagnetica rispetto a quella di vetro di spin. In altre parole, bisogna calcolare il diagramma di fase del sistema rispetto ai parametri del modello: per fare ciò, dobbiamo calcolare la stabilità dei minimi dell'energia libera. Concentriamoci prima sulla transizione paramagnetico-ferromagnetica: dobbiamo calcolare l'Hessiano dell'energia libera rispetto ai modi α, prendendo il limite $q = \bar{q} = m_\alpha = \bar{m}_\alpha = 0$.

La matrice ottenuta è

$$H_{\alpha\alpha} = \begin{bmatrix} w_\alpha & w_\alpha^2 \\ w_\alpha^2 & w_\alpha \end{bmatrix}$$

e quindi la fase paramagnetica diventa instabile quando il modo singolare più forte di $\boldsymbol{w}$ soddisfa

$$w_\alpha^2 > 1.$$

Alla stessa stregua possiamo studiare le altre transizioni e, complessivamente, troviamo il diagramma di

system might be trapped into a set of similar configurations. These mean-field equations are satisfied at the minima of the free energy of the system. We can now focus on the different phases of the system. Following the traditional analysis in spin glass theory, we can look for the stability of the following regions

- $m_\alpha = \bar{m}_\alpha = q = \bar{q} = 0$ **the paramagnetic phase**. It occurs typically at high temperature (for small σ) and weak couplings (small w_α)
- $m_\alpha, \bar{m}_\alpha, q, \bar{q} \neq 0$ **the ferromagnetic phase**. In this phase, we expect that the system will condensate over the singular modes: the equilibrium configurations will have a macroscopic overlap with one or many of these modes.
- $m_\alpha = \bar{m}_\alpha = 0$ and $q, \bar{q} \neq 0$ **the spin-glass phase**. In this phase, the system is stuck into few low energy configurations but these configurations are not correlated to the singular modes of $\boldsymbol{w}$

On order to recover configurations that are correlated to the signals w_α we need to enter the ferromagnetic phase. Intuitively, we can assume that when the noise matrix is weak (small values of σ) and that the signal is also weak (small w_α) we should be in the paramagnetic phase. The question is to understand what would trigger the ferromagnetic phase over the spin-glass one (if we wish to avoid the latter for instance). In other words, we want to compute the phase diagram of the system with respect to the parameters of the model. To do that, we need to compute the stability of the minima of the free energy. Let's focus first on the paramagnetic-ferromagnetic transition. We need to compute the Hessian of the free energy w.r.t. the modes α, taking the limit $q = \bar{q} = m_\alpha = \bar{m}_\alpha = 0$.

The obtained matrix is

$$H_{\alpha\alpha} = \begin{bmatrix} w_\alpha & w_\alpha^2 \\ w_\alpha^2 & w_\alpha \end{bmatrix}$$

and therefore the paramagnetic phase becomes unstable when the strongest singular mode of $\boldsymbol{w}$ satisfies

$$w_\alpha^2 > 1.$$

Similarly we can study the other transitions and, overall, we find the phase diagram reported in Figure 2. A

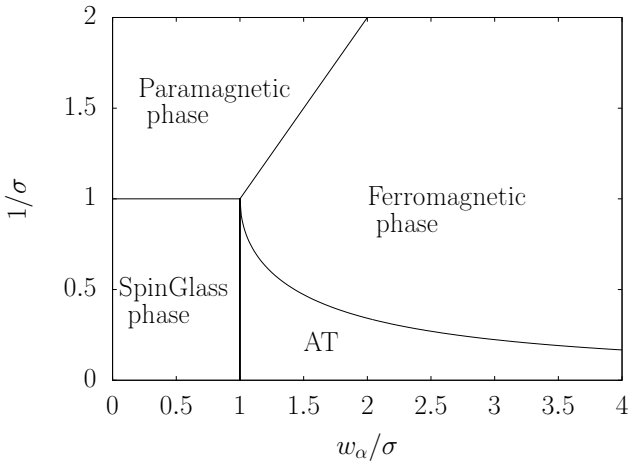

Figura 2: *Diagramma di fase della RBM in funzione del rumore σ e del modo singolare maggiore w_α. Aggiungiamo anche la linea-AT sotto la quale la soluzione simmetrica replica diventa instabile.*
Phase diagram of the RBM as a function of the noise σ and the highest singular mode w_α. We add the AT-line below which, the replica-symmetric solution becomes unstable.

fase riportato in Figura 2.

Qualche commento è d'obbligo: il diagramma di fase ci dice dove l'apprendimento deve guidare il sistema, aggiustando la matrice dei pesi, affinchè finisca in una fase ferromagnetica. La fase ferromagnetica, assumendo che i modi singolari appresi di $\boldsymbol{w}$ siano correlati al set di dati, descrive una fase in cui le configurazioni di equilibrio sono esse stesse correlate al set di dati. In particolare, è possibile descrivere più in dettaglio le proprietà della fase ferromagnetica come in [6]. A seconda della distribuzione usata per le matrici $\boldsymbol{u}$ e $\boldsymbol{v}$ possiamo avere o una fase dominata solo dal modo più forte oppure una fase in cui i minimi dell'energia libera sono costituiti da una composizione di modi.

Learning dynamics — come accennato prima, ci aspettiamo che il comportamento della RBM gaussiana "lineare" sia corretto all'inizio dell'apprendimento anche per la RBM binaria perché gli accoppiamenti sono deboli. Ciò può essere verificato eseguendo una piccola espansione del gradiente nell'accoppiamento e proiettandola sulla SVD di $\boldsymbol{w}$. Otteniamo, per i modi singolari

$$\frac{dw_\alpha}{dt} = w_\alpha \left[\langle \hat{s}_\alpha^2 \rangle - 1\right],$$

che corrispondono alle equazioni ottenute per la RBM gaussiana con $\sigma_v = \sigma_h = 1$. Questo conferma che, quando gli accoppiamenti sono deboli, i modi singolari più forti del set di dati attiveranno l'apprendimento, ancora una volta. Sperimentalmente si può

few comments are in order: the phase diagram tells us where the learning should bring the system, adjusting the weight matrix, in order to end up in a ferromagnetic phase. The ferromagnetic phase, assuming that the learned singular modes of $\boldsymbol{w}$ are correlated to the dataset, describes a phase where the equilibrium configurations are correlated to the dataset. In particular, it is possible to describe into more details the properties of the ferromagnetic phase as in [6]. Depending on the distribution used for the matrices $\boldsymbol{u}$ and $\boldsymbol{v}$ we can have either a phase dominated only by the strongest mode or a phase where the minima of the free energy is made of composition of modes.

Learning dynamics — as mentioned before, we expect that the behavior of the "linear" Gaussian RBM is correct at the beginning of the learning for the binary RBM because the couplings are weak. This can be verified by making a small coupling expansion of the gradient and projecting it on the SVD of $\boldsymbol{w}$. We obtain for the singular modes

$$\frac{dw_\alpha}{dt} = w_\alpha \left[\langle \hat{s}_\alpha^2 \rangle - 1\right],$$

which correspond to the equations obtained for the Gaussian-RBM with $\sigma_v = \sigma_h = 1$. It confirms that, when the couplings are weak, the strongest singular modes of the dataset will trigger the learning, once again. Experimentally it can be observed on a complex dataset that, during the first iterations of the

osservare su un dataset strutturato che, durante le prime iterazioni dell'apprendimento, le caratteristiche apprese dalla matrice dei pesi $\boldsymbol{w}$ sono infatti quasi indistinguibili dai modi principali del dataset. Tuttavia, dopo un lungo periodo di addestramento, queste cambiano completamente e sembrano avvicinarsi alle componenti ottenute in un'analisi delle componenti indipendenti [14]. Rimane da esplorare la formazione di *patterns* e la loro evoluzione attraverso le dinamiche di apprendimento, anche se esistono alcuni risultati nel caso di una RBM con uno o due nodi nascosti [15].

Conclusioni

Come visto in questo articolo, i modelli utilizzati nel machine learning possono essere molto vicini a quelli studiati dal fisico. Il caso discusso in questo articolo è particolare poiché la RBM corrisponde esattamente al modello Ising ed ha portato una nuova serie di problemi interessanti. Ad esempio, il diagramma di fase deve essere compreso in maggiore dettaglio. Forse ancora più interessante è il modo in cui le dinamiche di apprendimento guidano la rete da una fase completamente paramagnetica ad una regione che comprende una fase dove è possibile un *retrieval* compositivo. Durante questo processo, avviene la formazione di patterns non lineari, collegati al set di dati in esame, all'interno del sistema: resta da capire come emergono questi pattern e il loro legame con le proprietà statistiche del set di dati.

learning, the features learned by the weight matrix $\boldsymbol{w}$ are indeed almost indistinguishable from the principal modes of the dataset. However, after a long training time, they change completely and seems to get closer to the components obtained in an Independent Component Analysis[14]. The formation of patterns and their evolution via the learning dynamics remain to explore, even if some results exist in the case of an RBM with one or two hidden nodes[15].

Conclusion

As seen in this article, the models used in machine learning can be very close to the ones studied by physicist. The case discussed in this paper is particular since the RBM corresponds exactly to the Ising model, yet it brought a new set of interesting problems. For instance, the phase diagram remains to be understood into more details. Maybe even more interestingly is how the learning dynamics drives the system from a completely paramagnetic phase to a region involving a compositional retrieval phase. During this process, non-linear pattern formation occurs within the system and linked to the dataset under consideration: understanding how these patterns emerge and their link to the statistical properties of the dataset remain to be understood.

[1] Y. LeCun, Y. Bengio, G. Hinton.: *Deep Learning*, Nature, 521 (2015) 436.

[2] V. Mnih, et al., *Boltzmann machines: Constraint satisfaction networks that learn*, Nature, 518 (2015) 529.

[3] G. Hinton, et al.,: *Human-level control through deep reinforcement learning* Carnegie-Mellon University, Department of Computer Science, Pittsburgh, PA (1984).

[4] N. Le Roux, Y. Bengio, *Representational power of restricted Boltzmann machines and deep belief networks*, Neural Comp., 20 (2008) 1631.

[5] R. Karakida, M. Okada, S. Amari, *Analyzing feature extraction by contrastive divergence learning in RBMs*, in Deep learning and representation learning workshop: NIPS, (2014).

[6] A. Decelle, G. Fissore, C. Furtlehner, *Thermodynamics of restricted Boltzmann machines and related learning dynamics*, J. Stat. Phys., 172 (2018) 1576.

[7] D.JC. MacKay, D.JC. Mac Kay *Information theory, inference and learning algorithms*, Cambridge Univ. Press., Cambridge (UK) (2003).

[8] C.M. Bishop, *Pattern recognition and machine learning*, Springer Press, Berlin (2006).

[9] M. Mézard, G. Parisi, M.A. Virasoro, *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*,, World Sci. Publ., Singapore (1987).

[10] J Tubiana, R. Monasson, *Emergence of compositional representations in restricted Boltzmann machines*, Phys. Rev. Lett. 118 (2017) 138301.

[11] A. Barra, et al., *Phase diagram of restricted Boltzmann machines and generalized Hopfield networks with arbitrary priors*, Phys. Rev. E 97 (2018) 022310.

[12] R.L. Stratonovich, *On a method of calculating quantum distribution functions*, Soviet Physics Doklady, 2 (1957) 416.

[13] J. Hubbard, *Calculation of partition functions*, Phys. Rev. Lett., 3 (1959) 77.

[14] A. Hyvärinen, E. Oja, *Independent component analysis: algorithms and applications*, Neural Net., 13 (2000) 411.

[15] M.E. Harsh, et al., *'Place-cell'emergence and learning of invariant data with restricted Boltzmann machines*, J. Phys. A., 53 (2020) 174002.

Aurélien Decelle: è ricercatore in Meccanica Statistica presso l'Universidad Complutense de Madrid. I suoi interessi di ricerca ruotano attorno ai sistemi disordinati, con particolare attenzione a problemi di inferenza e machine learning.

Aurélien Decelle: is a researcher in statistical physics at the Universidad Complutense de Madrid. His research focuses on disordered systems, with a particular interest in inference problems and machine learning.