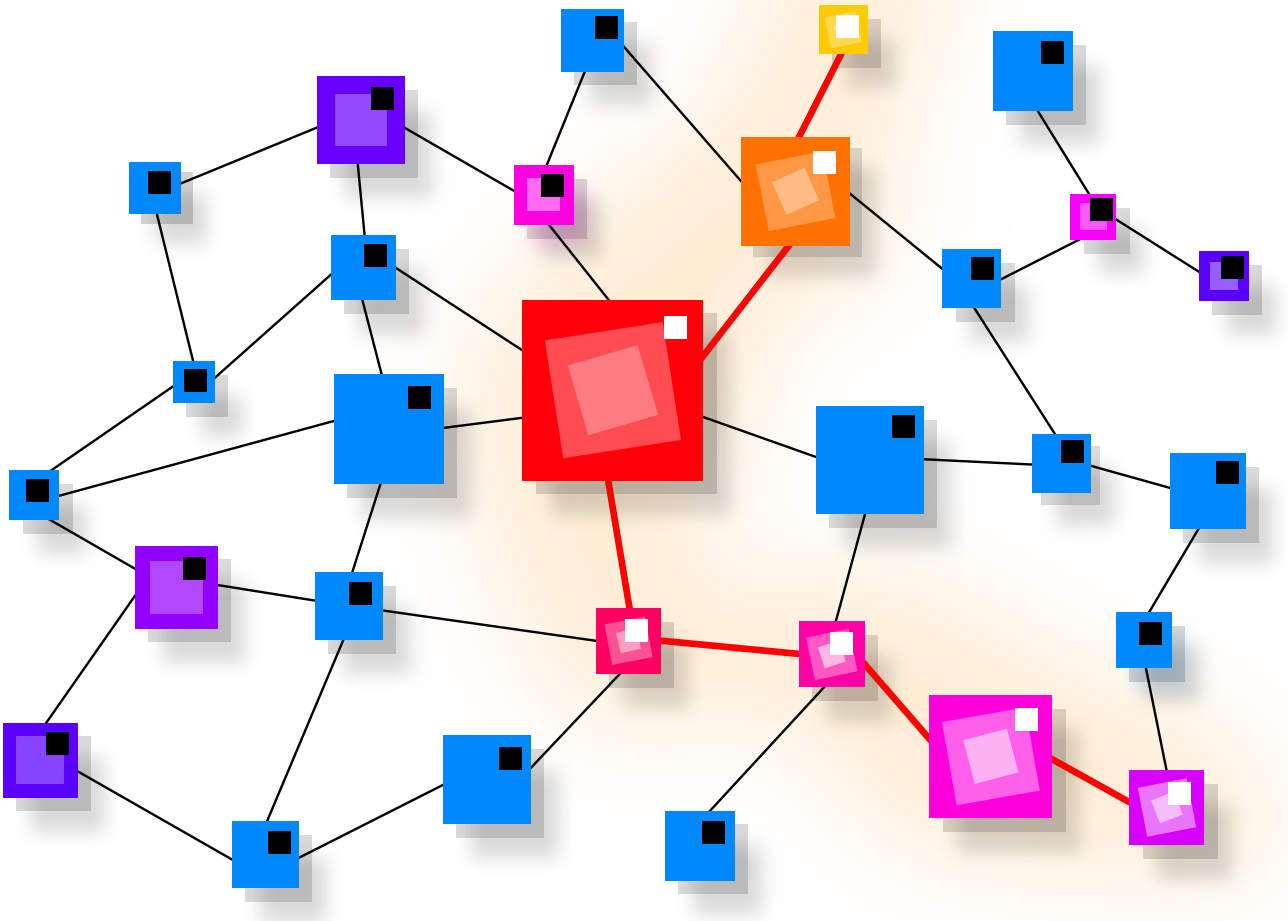
Numero XVI
Anno 2020

DIPARTIMENTO DI MATEMATICA E FISICA
"Ennio De Giorgi"

UNIVERSITÀ
DEL SALENTO

Viaggio nella Scienza

Ithaca



Intelligenza artificiale

Ithaca: Viaggio nella Scienza

Una pubblicazione del Dipartimento di Matematica e Fisica "Ennio De Giorgi" dell'Università del Salento.

Registrazione presso il Tribunale di Lecce n. 6 del 30 Aprile 2013.
e-ISSN: 2282-8079

Ithaca: Viaggio nella Scienza
è disponibile sul sito:

http://ithaca.unisalento.it/

Scriveteci all'indirizzo:

ithaca@unisalento.it

Ithaca
Viaggio nella Scienza

XVI 2020

In questo numero

Il tema di questo numero è l'Intelligenza Artificiale: mentre questa dilaga nelle nostre case e nel nostro modo di lavorare, cambiando volto alla società, molto rimane ancora da capire sul modo in cui le sue reti neurali artificiali di fatto operano. Proprio per riuscire ad orientarsi in questo mare magno e supplire quindi al lettore una prospettiva teorica con cui guardare questi progressi tecnologici, il *leitmotiv* del numero, pensato per un'*audience* di fisici e matematici, è la meccanica statistica dei sistemi complessi (alla quale è dedicata la lezione mancata, posta in coda al presente numero, al fine di agevolare il lettore nella comprensione dei vari contributi).

Il numero XVI di Ithaca ha un'Alpha ed un'Omega *istituzionali* per il soggetto del volume: l'Alpha è scritto da Paolo Del Giudice e Maurizio Mattia, dell'Istituto Superiore di Sanità, che ci mostrano l'impressionante crescita culturale nella modellistica delle reti neurali biologiche; di contro Giorgio Buttazzo, della Scuola Superiore Sant'Anna, scrive l'Omega, riassumendo tale crescita nelle reti neurali artificiali: dalla biologia all'ingegneria il percorso del numero attraverserà fisica e matematica.
Nello specifico Paolo Del Giudice e Maurizio Mattia amalgamano storia e scienza, lasciandoci addentrare in problemi via via più complessi, dalla rumorosa decisione di un neurone sull'emettere (o meno) un segnale elettrico, alla ben più complessa decisione del nostro cervello sul se frenare (o meno) davanti ad un apparente semaforo rosso, lasciandoci intuire come la cognizione si sviluppi su una gerarchia di scale, spaziali e temporali.
A seguire, l'articolo di Elena Agliari ed Adriano Barra fa da gradiente dalle reti neurali biologiche a quelle artificiali: iniziando con modelli ispirati alla biologia, si approda a modelli paradigmatici nel machine learning al fine di affrontare il fenomeno della cognizione nelle reti neurali in astratto (alla volta di un armonioso connubio tra l'apprendimento e successivo impiego dell'informazione appresa). Prosegue sulla stessa linea Daniele Tantari che, sfruttando una metafora tra il modo con cui uno studente impara da un docente e l'apprendimento supervisionato delle macchine moderne, ripercorre il lungo cammino dell'apprendimento automatico, mostrandoci come anche il concetto stesso di imparare stia evolvendo per esse. A questo contributo segue quello di Aurélienne Decelle, il quale analizza in dettaglio un archetipo del *machine learning* particolarmente caro a chi investiga nelle scienze dure: la macchina di Boltzmann. Nel suo articolo il lettore viene portato per mano attraverso un susseguirsi di varianti sul tema via via più complesse e, conseguentemente, più capaci, al fine di mostrare in cosa consti operativamente l'apprendimento di una macchina.
La Boltzmann machine è una macchina relativamente piccola, di contro molte architetture odierne - in particolare quelle che hanno dato vita alla rivoluzione del deep learning - sono telai di molti strati di neuroni connessi a cascata: di queste ci parla prima Guido Sanguinetti, che ne sviscera i talloni d'Achille e pone enfasi sull'importanza di un controllo tanto sull'incertezza associata alle loro predizioni quanto sull'interpretabilità stessa del loro operato portandoci così ad interrogarci sull'etica inerentemente l'impiego stesso del machine learning nella società. Il testimone passa poi a Carlo Lucibello il quale impiega concetti e tecniche in auge nella meccanica statistica dei sistemi complessi, focalizzandosi sul percettrone binario e sul percettrone a

molti strati, sollevando la delicata questione della scelta delle variabili (e.g. pesi sinaptici analogici vs digitali), e, di fatto, mostrando quanto questa scelta possa influire sulla dinamica di queste reti.
Nel vivo della questione si colloca a seguire il contributo di Matteo Marsili che pone accento sulla cruciale questione della rilevanza nell'apprendimento statistico e mette in rilievo differenze strutturali tra sistemi fisici di classico appannaggio della meccanica statistica e reti neurali artificiali, spostando il *focus* dalla meccanica statistica alla teoria statistica dell'apprendimento ed evidenziando le prime discrepanze tra il regime d'azione della statistica classica e quello del machine learning. Naturale conseguenza di questo è il successivo contributo di Jean Barbier sull'inferenza ad alta dimensionalità: un preziosissimo Bignami di questa disciplina che sta nascendo all'intersezione di meccanica statistica, inferenza statistica, teoria dell'informazione e complessità algoritmica e che naturalmente si candida ad essere uno dei pilastri sui quali erigere una teoria per l'intelligenza artificiale. Sempre tenendo il *focus* sull'informazione, o sull'entropia a secondo della propria preferenza, si innesta in maniera naturale il contributo di Michele Castellana, il quale rivede –criticamente e sistematicamente– il principio di massima entropia in chiave inferenziale, ponendo l'enfasi sull'importanza dell'errore di misura nell'impiego di questi metodi e chiudendo le disquisizioni canoniche di fisici e matematici.

A seguire, Valerio Basile ci parla dei dietrole-quinte di una delle applicazioni più importanti dell'Intelligenza Artificiale: la traduzione automatica, punta d'iceberg di branche prolifiche quali la linguistica computazionale ed il natural language processing, mentre Dino Bachas e Stefania Spagnolo si concentrano su un'altra parimenti cruciale applicazione, a dire come il *machine learning* stia cambiando il modo di setacciare i dati nella fisica delle alte energie. L'ultima doverosa applicazione viene passata in rassegna da Alfredo Braunstein, Luca Dall'Asta e Alessandro Ingrosso, i quali mostrano come queste tecniche siano oltremodo benvenute, per non dire assolutamente necessarie, negli attuali schemi di tracciamento automatico, ed in generale per il contenimento epidemico, inerentemente i casi di infezione da Covid-19.

Il volume si chiude riportando l'intelligenza artificiale nel suo dominio classico di pertinenza, l'ingegneria informatica, con Giorgio Buttazzo che ci offre una ricca carrellata dei modelli in auge in questa disciplina, dal neurone artificiale di McCulloch & Pitts della metà del secolo scorso alle varie architetture e tecniche di impiego odierno, chiudendo il volume rimarcando l'interrogativo etico sul rapporto uomo macchina e la sua (veloce) evoluzione.

Buona lettura,
il Comitato di Redazione

Tanto rumore per nulla? Il ruolo delle fluttuazioni nella dinamica delle reti nervose

Quell'altro organo che chiamiamo cervello, quello con cui veniamo al mondo, quello che trasportiamo nel cranio e che trasporta noi affinché noi trasportiamo lui, non è mai riuscito a produrre altro che intenzioni vaghe, generiche, diffuse, e soprattutto poco variate, riguardo a ciò che le mani e le dita dovranno fare.

"*La Caverna*", José Saramago

Paolo Del Giudice
Centro Nazionale per la Protezione dalle Radiazioni e Fisica Computazionale, Istituto Superiore di Sanità, Roma

Maurizio Mattia
Centro Nazionale per la Protezione dalle Radiazioni e Fisica Computazionale, Istituto Superiore di Sanità, Roma

L'idea che i sistemi di Intelligenza Artificiale (IA) possano trarre vantaggiosa ispirazione dalle scienze del cervello è molto naturale, ha permeato in varie forme gli sviluppi di IA per decenni e, nella sua formulazione generica, costituisce ormai un luogo comune. Non sorprende quindi che in un numero della rivista dedicato alla IA trovi posto un articolo dedicato ad alcuni aspetti della modellistica teorica del sistema nervoso. Se però si va a scandagliare la storia dei rapporti tra IA e teorie del cervello (che è complessa e non tenteremo di riassumerla) si scopre che l'ispirazione di cui sopra ha preso forme qualitativamente diverse; ne discuteremo brevemente in una parte introduttiva per fornire il contesto di quanto diremo in seguito. Passeremo poi a descrivere alcuni modelli fisico-matematici di attività nervosa a varie scale, mettendo a fuoco il ruolo delle fluttuazioni casuali come elemento costitutivo della dinamica neuronale.

Pur accettando per semplicità il punto di vista, piuttosto diffuso, secondo il quale il cervello effettua delle *computazioni* sugli input ambientali, e ove opportuno ne traduce il risultato in *azio-*

ni (trascurando cioè la natura profondamente *relazionale* del loop percezione-azione), in un approccio ad un sistema intelligente ispirato alla conoscenza del cervello si può ad esempio: *i)* emulare, in qualunque forma conveniente, alcune elaborazioni già identificate nel sistema nervoso (imitare l'analisi in frequenza effettuata dalla coclea sul suono; riprodurre parte della gerarchia di filtri implementati nelle aree visive, come quelli suggeriti dalla *independent component analysis*; ricostruire - sulla base di quanto se ne è finora capito in neurofisiologia - la pianificazione e l'esecuzione del movimento per sviluppare robot con capacità motorie complesse, ecc.); *ii)* sviluppare sistemi di apprendimento automatico, i cui elementi siano più o meno vagamente ispirati alle unità di elaborazione del sistema nervoso, e la cui architettura venga determinata da un processo di ottimizzazione rispetto alle prestazioni su specifici compiti (*reti neurali*, *machine learning*); *iii)* sviluppare, in diretta connessione con i dati sperimentali, modelli fisico-matematici in grado di cogliere elementi fondamentali della dinamica di popolazioni di cellule nervose fortemente interconnesse, per comprendere in che modo le loro proprietà di risposta agli stimoli e di auto-organizzazione possano generare delle *primitive computazionali* per lo svolgimento di compiti cognitivi complessi.

È di quest'ultimo approccio che tratteremo nel seguito.

Alcuni principi dinamici

Un principio molto generale di organizzazione dei sistemi biologici è la ricerca di un compromesso ottimale tra *exploitation* e *exploration*. Da un lato un organismo deve poter fornire risposte riproducibili e funzionalmente adeguate, determinate da adattamenti alle richieste ambientali. Dall'altro è vantaggioso che esso non rimanga fossilizzato nelle soluzioni dettate dalle situazioni contingenti cui è stato esposto, ed elabori una strategia di esplorazione del proprio spazio degli stati, in grado di generare soluzioni non accessibili attraverso un'ottimizzazione 'locale'. Le direzioni di questa esplorazione non dovrebbero essere troppo vincolate dalla configurazione corrente, e questo ci fa già intuire il possibile ruolo costruttivo di una componente di *rumore*.

L'esempio forse più familiare di questo principio generale è costituito dall'evoluzione biologica, prodotta della combinazione di variabilità casuale del patrimonio genetico e di pressione selettiva per la sopravvivenza del più adatto ('caso e necessità'), in un ambiente che in generale è mutevole nel tempo.

Nei classici modelli matematici di evoluzione la pressione selettiva è descritta da un "paesaggio di fitness" (eventualmente dinamico), e l'equilibrio *exploitation/exploration* si esprime nella tendenza a popolare le configurazioni di massima fitness (i picchi locali del paesaggio), come risposta ai vincoli ambientali contingenti, e contemporaneamente nella esplorazione del paesaggio attraverso variazioni casuali del genoma, che rendono accessibili 'soluzioni' evolutive che non sarebbero localmente accessibili.

Cambiando prospettiva dall'evoluzione delle specie allo sviluppo embrionale del singolo individuo, fin dagli anni '40 fu formulata la metafora di un 'paesaggio epigenetico', del quale lo sviluppo cellulare dell'embrione segue le pendenze e le biforcazioni, che rappresentano in modo astratto un sistema di vincoli per lo sviluppo.

Vedremo nel seguito che simili schematizzazioni teoriche, basate su un 'paesaggio' come rappresentazione di un sistema di vincoli, e su fluttuazioni casuali come opportunità di esplorazione dello spazio degli stati, hanno un corrispettivo diretto nella modellizzazione del sistema nervoso.

Il sistema nervoso deve in effetti risolvere un problema di exploitation/exploration analogo: deve da un lato assicurare l'affidabilità del circuito percezione-azione nella sue varie forme, e la robustezza delle rappresentazioni che costruisce per classificare la realtà; se così non fosse non sarebbe neanche possibile la concordanza, che ci appare invece naturale, tra ciò che diversi cervelli classificano come 'rosso' o 'cane', ed i movimenti dei nostri simili nel reagire agli stimoli ambientali ci apparirebbero impredicibili. D'altra parte, non solo diversi cervelli possono costruire sequenze di pensieri ed immagini mentali profondamente diverse a parità di condizioni ambientali, com'è ovvio dall'esperienza comune, ma in determinate condizioni, ed in modo misurabile, l'attività nervosa varia in modo apparentemente casuale in risposta agli stessi stimo-

li, e (ne vedremo un esempio) può alternare in modo stocastico tra le possibili rappresentazioni mentali alternative dello stesso stimolo.

Se i picchi del paesaggio di fitness dell'evoluzione biologica rappresentano soluzioni evolutive 'localmente ottimali', i picchi di un analogo paesaggio che funga da palcoscenico per la dinamica nervosa possono esprimere ad esempio l'ottimalità locale delle rappresentazioni neuronali di scene visive o di atti motori, a seconda del contesto.

In questa metafora, il cervello transisce continuamente tra gli stati che corrispondono a tali rappresentazioni, in parte a causa di input mutevoli, in parte sotto la spinta di elementi dinamici endogeni. Come vedremo, la disposizione a queste transizioni si deve sia alla capacità del sistema di modificare la porzione del paesaggio in cui temporaneamente risiede, sia a diverse sorgenti di rumore.

Notiamo di passaggio che se, al limite, descriviamo la dinamica delle rappresentazioni espresse nel cervello come transizioni quasi-discrete tra configurazioni localmente ottimali, di fatto ci ricolleghiamo in parte ad una antica diatriba sulla natura continua o discreta della percezione cosciente, che risale a D. Hume e passa per W. James e H. Bergson. È comunque corretto dire che, a livello sia sperimentale che teorico, il dibattito sulla possibilità di interpretare la dinamica nervosa in termini di transizioni discrete tra stati ben definiti è tuttora aperto.

Assumendo di adottare questo quadro concettuale di riferimento per la descrizione della dinamica del sistema nervoso, dobbiamo definire quali quantità sperimentalmente osservabili definiscano gli stati del sistema e che cosa, e come, determini il 'paesaggio' in cui tali stati evolvono.

A prima vista, l'eterogeneità delle nostre esperienze soggettive in relazione a diversi stati mentali sembra opporsi alla formulazione di un'astrazione di valore generale. In effetti, nella nostra esperienza soggettiva, le diverse situazioni concrete nelle quali si istanzia la tensione tra ottimalità locale ed esplorazione casuale a cui abbiamo accennato ci appaiono qualitativamente diverse: riconoscere un sapore, classificare una forma, memorizzare una informazione, decidere un movimento, corrispondono a rapporti con il nostro corpo, nelle sue funzioni sensoriali, cognitive o motorie, che avvertiamo come appartenenti a domini diversi. Ma appena non appena ci allontaniamo dall'involucro sensoriale o motorio del sistema nervoso (cioè dalle sue componenti direttamente coinvolte nella trasduzione degli stimoli o nelle istruzioni motorie) queste distinzioni qualitative sfumano in una trama complessa di messaggi stereotipati (potenziali d'azione o 'spike') scambiati tra elementi sostanzialmente molto simili tra loro (neuroni) attraverso giunzioni (sinapsi) che si presentano in pochi tipi largamente distribuiti in tutto il cervello. Se un agente microscopico potesse intercettare gli spike scambiati localmente in qualunque area del cervello non avrebbe modo di qualificarli come 'spike visivi', 'uditivi', 'motori' o altro: solo l'organizzazione spazio-temporale del flusso di spike scambiati definisce la 'computazione'.

Questo fatto (che costituisce una scelta evolutiva peculiare e non ovvia) ci fornisce dunque un livello di descrizione di appropriata generalità e ci permette di definire lo stato del sistema come la collezione degli spike scambiati tra i neuroni di una popolazione di interesse ad un certo istante. Inoltre, per dar conto della componente stocastica di 'exploration', i processi di generazione e scambio degli spike includeranno componenti di 'rumore', di cui esamineremo nel seguito il senso e l'origine. Notiamo subito che questa definizione dello stato di una popolazione neuronale è in certo senso una definizione 'microscopica', analoga a quella che definisce lo stato di un gas ad un certo istante come l'insieme delle posizioni e velocità delle sue molecole a quell'istante. In quell'ambito (in cui il 'rumore' è determinato dalla temperatura), è molto utile adottare una descrizione in termini delle distribuzione di probabilità di quantità osservabili (meccanica statistica), e da questa derivare leggi macroscopiche (termodinamica). Vedremo sommariamente che, in qualche misura, lo studio teorico della dinamica neuronale procede attraverso un percorso simile.

Riguardo a cosa e come si determini la forma del 'paesaggio' in cui il sistema neuronale evolve, l'idea chiave, di nuovo in analogia con modelli familiari in fisica, è che essa sia primariamente legata alle interazioni tra i neuroni, cioè alla forza delle sinapsi che mediano lo scambio di spike tra coppie di neuroni. Come abbiamo accennato,

la forma del paesaggio esprime un insieme di vincoli a cui il sistema deve adattarsi in modo ottimale, identificando ad esempio, con i suoi picchi (o valli, secondo le convenzioni adottate) gli stati neuronali che forniscono la risposta ottimale ad una classe di stimoli. La forma del paesaggio deve quindi riflettere il modo in cui, nelle sue relazioni con il mondo esterno e con altre popolazioni neuronali, la popolazione in questione ha strutturato nel tempo le interazioni tra i suoi neuroni in modo da scolpire i profili di picchi e valli per poi guidare la dinamica neuronale in modo appropriato alle richieste ambientali e comportamentali.

Vediamo quindi che siamo di fronte a dinamiche accoppiate su scale di tempo diverse: quella 'veloce' della dinamica neuronale, sulla cui scala temporale il paesaggio è quasi statico (a meno di effetti locali cui abbiamo accennato e che vedremo in seguito), ed una dinamica 'lenta' che scolpisce il paesaggio in funzione dei vincoli esterni ed interni agendo anche sugli accoppiamenti tra i neuroni; quest'ultima viene spesso identificata, in varie forme, come 'apprendimento'; ne accenneremo soltanto, ma è una parte importante dello schema teorico.

Formalizzare gli elementi di base: neuroni e sinapsi

Nella formulazione di modelli fisico-matematici di sistemi biologici si accetta comunemente un certo livello (spesso elevato) di semplificazione e schematizzazione del dato biologico: questo è il prezzo che si accetta di pagare per poter rendere i modelli trattabili formalmente (o anche solo realisticamente simulabili al calcolatore). Tale semplificazione viene a volte motivata, soprattutto dai fisici, non solo come una inevitabile limitazione, ma come l'espressione del fatto (imparato dalla meccanica statistica) che spesso il comportamento collettivo di grandi insiemi di unità interagenti non dipende troppo dai dettagli delle singole unità. Come in altri domini scientifici, il problema difficile è identificare il livello appropriato di semplificazione in grado di assicurare i vantaggi di cui sopra, senza però allontanare troppo il modello dalla controparte biologica che intende descrivere.

Nel caso dei modelli dinamici di neurone (non è così per quelli sinaptici) si è riusciti a costruire una gerarchia di modelli di semplicità crescente, partendo da modelli 'realistici' in grado di confrontarsi quantitativamente con misure sperimentali dettagliate e passando attraverso una serie di approssimazioni controllate, in modo tale che ad ogni passo di semplificazione è abbastanza chiaro l'impatto dei dettagli che si è scelto di ignorare o accorpare.

La caratteristica comune a tutti i modelli neuronali consiste nel descrivere sistemi *eccitabili*: informalmente, sistemi che rispondono in modo quasi lineare ad input piccoli, ma che per input superiori ad una soglia esibiscono una risposta molto nonlineare, ed essenzialmente stereotipata. L'eccitabilità cellulare non è esclusiva dei neuroni (si pensi alle cellule cardiache). Nel sistema nervoso, le proprietà di risposta nonlineare sono la chiave per comprendere le capacità computazionali dei neuroni.

Non tentiamo di riassumere la gerarchia di modelli neuronali, e le strategie di semplificazione ed analisi che ne sono alla base; adotteremo piuttosto direttamente una classe di modelli che ha avuto un ruolo chiave nella modellistica teorica (e che si colloca verso il fondo della sequenza di semplificazioni a cui abbiamo accennato): il neurone *integrate-and-fire* (IF).

Mentre in modelli più complessi le proprietà eccitabili del neurone sono esplicitamente descritte in termini della dinamica di gradi di libertà interni al neurone, nel modello IF il neurone è descritto da una sola variabile di stato, il potenziale di membrana $V(t)$ (la cui variazione nei modelli dettagliati determina l'apertura o chiusura di canali ionici sulla membrana del neurone, che a loro volta ne determinano il comportamento nonlineare) la cui dinamica lineare guidata dall'input sinaptico $I(t)$ è

$$\frac{dV}{dt} = -\frac{V(t)}{\tau} + I(t)\,.$$

Le proprietà nonlineari necessarie per descrivere la generazione di uno spike (espressione della risposta stereotipata sopra soglia del sistema eccitabile, cui facevamo riferimento) sono condensate matematicamente in una condizione al contorno di questa equazione per il potenziale di membrana: quando V supera una soglia θ,

lo spike viene emesso, V viene riportato ad un valore di riferimento H, per riprendere dopo un periodo 'refrattario assoluto' τ_0 la sua evoluzione in regime lineare, fino al prossimo eventuale raggiungimento della soglia. In assenza di input, $V(t)$ ha un decadimento esponenziale con tempo caratteristico τ. L'input sinaptico è determinato dalla combinazione lineare degli spike emessi dai K neuroni presinaptici:

$$I(t) = I_{\text{ext}}(t) + \sum_{j=1}^{K} J_j \sum_k \delta(t - t_{j,k} - d_j)\,.$$

Il k-esimo spike emesso dal neurone j-esimo al tempo t_j, k viene ricevuto con un ritardo assonale d_j ed è rappresentato da una δ di Dirac, e induce un salto del potenziale V pari all'efficacia sinaptica J_j. Il contributo di corrente sinaptica dovuto a neuroni esterni alla rete descritta, è riassunto nel termine I_{ext}.

Dai singoli neuroni alle reti: modularità, teoria di campo medio e descrizione mesoscopica

Come già accennato, le reti nervose mostrano una complessa organizzazione sia da un punto di vista strutturale che funzionale. Diverse componenti, che includono ad un livello macroscopico le aree corticali identificate più di un secolo fa dal neurologo tedesco Korbinian Brodmann, sono tra loro gerarchicamente connesse contribuendo con differenti ruoli all'emergenza delle funzioni cognitive. L'eterogeneità di questi elementi si manifesta principalmente ad una scala spaziale "mesoscopica" più ridotta, dove la diversa citoarchittura appare evidente nella composizione cellulare dei cosiddetti *strati corticali* (*layers*) (Figura 1a). Ogni strato in queste reti strutturate contribuisce con un ruolo specifico alla elaborazione e al trasferimento dell'informazione ricevuta da altre regioni cerebrali dando luogo ad un circuito "canonico" che si ripete quasi invariato come le piastrelle di un pavimento. In questo circuito l'input sensoriale da strutture subcorticali come il talamo viene acquisito dalle sottoreti del quarto strato, amplificato e modulato da quelle del quinto e ritrasmesso dai neuroni degli strati più esterni. Questo modulo mesoscopico è riconoscibile come un fascio di fibre perpendicolare alla superficie corticale con una sezione di un centinaio di micron identificata alla fine degli anni '50 del secolo scorso dal neurofisiologo statunitense Vernon B. Mountcastle (Figura 1b). Le reti nervose condividono questo principio costituente di modularità con altri sistemi biologici descrivibili come reti, quali le cosiddette vie metaboliche e la regolazione genica, per citare un paio di esempi. La modularità infatti rende un sistema più versatile e quindi vantaggioso da un punto di vista evolutivo. Rende ad esempio più semplice la riorganizzazione delle connessioni tra unità funzionali, in confronto al costo richiesto nel trasformare la matassa inestricabile di legami associata ad una rete omogenea.

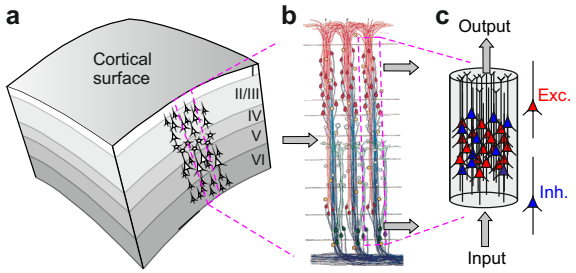


Figura 1: *Rappresentazione schematica dell'organizzazione a strati della corteccia cerebrale (a) composta da moduli di fibre nervose segregate in "minicolonne" (b). I diversi strati corticali svolgono ruoli diversi tra cui la ricezione e l'invio di messaggi da e verso altre aree cerebrali e colonne (frecce grigie). Nella modellistica delle reti neuronali, una minicolonna è rappresentata da una popolazione di neuroni eccitatori ed inibitori (c). Pannello (b) adattato da [1].*

La modularità gerarchica delle reti cerebrali offre d'altro canto un'opportunità di semplificazione nella descrizione teorica delle loro dinamiche collettive. È infatti possibile immaginare di poter trascurare il dettaglio microscopico di queste colonne corticali, quale ad esempio la posizione e l'estensione spaziale delle cellule nervose che le compongono, per concentrarci principalmente sulle caratteristiche di "grana grossa" del loro stato, verificando a posteriori l'accuratezza di questa approssimazione (approccio, questo, ampiamente adottato in settori della fisica teorica ben sedimentati come quello della meccanica statistica).

In questo contesto, consideriamo ragionevolmente adeguato descrivere una colonna corticale come composta da neuroni semplificati del tipo *IF* introdotto precedentemente (Figura 1c). Per mantenere una sufficiente aderenza alla realtà biologica i neuroni di questo modulo corticale devono essere suddivisi in due categorie: eccitatori ed inibitori. Gli spike emessi dai primi inducono attraverso una trasmissione sinaptica *glutammatergica* (cioè mediata dal glutammato) un incremento del potenziale di membrane V nei neuroni post-sinaptici (neuroni bersaglio) favorendone l'avvicinamento alla soglia di emissione θ. Gli inibitori sono invece caratterizzati da una trasmissione sinaptica *GABA-ergica* (cioè mediata dal neurotrasmettitore GABA) che "iperpolarizza" il potenziale V del neurone bersaglio allontanandolo da θ. L'entità di questa deflessione dipende dalla cosiddetta *efficacia sinaptica* J che sarà positiva per i neuroni eccitatori e negativa per gli inibitori.

In corteccia cerebrale, ogni neurone riceve decine di migliaia di contatti sinaptici da neuroni che solo in parte appartengo al modulo corticale in questione. Il contributo maggioritario di corrente sinaptica dovuto agli spike emessi da neuroni "esterni" alla colonna corticale, può in prima approssimazione essere considerato come un *bagno termico* in analogia con la termodinamica, intendendo che questa componente si assume indipendente dall'evoluzione temporale della popolazione di neuroni "interni".

Altra caratteristica chiave dell'attività corticale è l'elevata irregolarità con la quale i neuroni emettono gli impulsi nervosi. A prima vista questi possono apparire come eventi casuali che si manifestano con frequenza variabile nel tempo. L'idea che un elaboratore di informazioni quale è il nostro cervello sia costituito da elementi dalla natura così rumorosa e quindi inaffidabile, può apparire quantomeno bizzarro. In realtà, questa apparente contraddizione svanisce se si considera l'informazione come distribuita tra i neuroni di una rete nervosa. In un elaboratore parallelo di siffatta natura, il singolo elemento darà un contributo molto parziale alla costituzione del messaggio da trasmettere o alla computazione da svolgere. Nelle reti nervose questa strategia è garantita dall'elevato grado di parallelismo implementato, ovverosia dal gran numero di elementi con cui ogni singolo neurone è in contatto attraverso le sinapsi del proprio albero dendritico. I grandi numeri in gioco rendono applicabile il teorema del limite centrale, portando ad avere correnti sinaptiche dominate dalle proprietà medie degli impulsi emessi dall'intera popolazione neuronale (naturalmente l'applicabilità del teorema richiede che le fluttuazioni dei molti input sinaptici siano anche tra loro indipendenti; come discuteremo brevemente appresso, a parte alcune condizioni patologiche la validità di questa ipotesi è in realtà abbastanza ben verificata nelle tipiche condizioni corticali).

Una tale codifica istantanea dell'informazione distribuita nello 'spazio', cioè tra i neuroni della rete, offre notevoli vantaggi come la resilienza ad eventuali malfunzionamenti dei singoli neuroni. Non solo: il singolo neurone potrà essere una unità di elaborazione relativamente lenta e inaffidabile, producendo messaggi/impulsi a frequenza ridotta ed in modo irregolare nel tempo con un conseguente risparmio di energia metabolica. Il prezzo da pagare nella scelta di questa strategia architetturale risiede nella necessità di dover costituire una fitta rete di comunicazione, difficoltà che però le reti cerebrali hanno risolto con successo.

Ma cosa determina l'irregolarità dei treni d'impulsi emessi da un neurone? Una corrente sinaptica $I(t)$ interamente guidata dalla sola media statistica degli impulsi ricevuti istante per istante sull'albero dendritico porterebbe ad una evoluzione del potenziale di membrana somatico $V(t)$ prettamente deterministica. Questo in realtà non accade, e il motivo risiede principalmente nel fatto che l'input sinaptico è una combinazione di contributi sia eccitatori che inibitori. Per capire le ragioni di questa affermazione, occorre prima considerare che i treni d'impulsi dei neuroni presinaptici con buona approssimazione si compongono linearmente sull'albero dendritico. La corrente $I(t)$ può allora essere vista come dovuta ad un treno di impulsi con statistica di Poisson. Combinando infatti un elevato numero di processi puntali (le sequenze di impulsi emessi dai singoli neuroni) con qualsivoglia statistica, il treno di impulsi risultante dalla loro somma è assimilabile ad un processo di Poisson con frequenza pari al prodotto $K\,\nu(t)$ del numero di neuroni presinaptici e la loro frequenza media di emissione

(Figura 2a). Questa è la conseguenza del teorema del limite centrale di Palm–Khintchin per i processi di 'renewal' stazionari generalizzato successivamente da B. Grigelionis al caso in cui la statistica di questi treni non sia omogenea nel tempo. Nell'ipotesi che i tempi di emissione degli spike tra due neuroni possano essere considerati indipendenti, l'intera sequenza di spike presinaptici sarà Poissoniana nel limite di frequenza $K\,\nu(t)$ infinita. Per quanto accennato in precedenza, ambedue queste condizioni sono ben soddisfatte dai neuroni corticali, sia per l'elevato numero di contatti sinaptici K, che per la ridotta frequenza $\nu(t)$ di 'sparo' (*firing*) registrata *in vivo* nelle cellule nervose. Quest'ultimo aspetto contribuisce a diluire nel tempo l'interazione tra due neuroni sinapticamente connessi, rendendoli nella pratica reciprocamente indipendenti.

Il continuo bombardamento di spike inibitori ed eccitatori se opportunamente bilanciati può portare ad avere ampie fluttuazioni della corrente presinaptica anche nel limite $K\,\nu \to \infty$. Per illustrare questo concetto è sufficiente derivare la media μ e la varianza σ^2 di questa corrente che per la sua natura Poissoniana risultano essere una combinazione lineare delle frequenza totali $K_I\nu_E(t)$ e $K_I\nu_I(t)$ come illustrato da Daniel J. Amit e Nicolas Brunel in [3]:

$$\begin{aligned} \mu(t) &= J_E K_E \nu_E(t) - J_I K_I \nu_I(t) \\ \sigma^2(t) &= J_E^2 K_E \nu_E(t) + J_I^2 K_I \nu_I(t)\,, \end{aligned}$$

in cui J_E e J_I rappresentano le intensità delle efficacie sinaptiche, cioè la variazione assoluta del potenziale V in corrispondenza all'arrivo di uno spike eccitatorio o inibitorio, rispettivamente. Queste due componenti si sottraggono nella media ma si sommano nella varianza e, a patto di riscalare le efficacie sinaptiche J_E e J_I, è sempre possibile ottenere dei valori finiti di μ e σ nel limite suddetto. Se la componente media della corrente non è sufficiente a portare il potenziale di membrana ad attraversare la soglia di emissione, gli spike verranno prodotti solo in presenza di fluttuazioni positive di $V(t)$. Questo porta al comportamento illustrato in Figura 2b, dove il potenziale di membrana del neurone spende la maggior parte del tempo a fluttuare al di sotto della soglia θ, emettendo spike a tempi irregolari. Come illustrato poco più di vent'anni fa da Carl

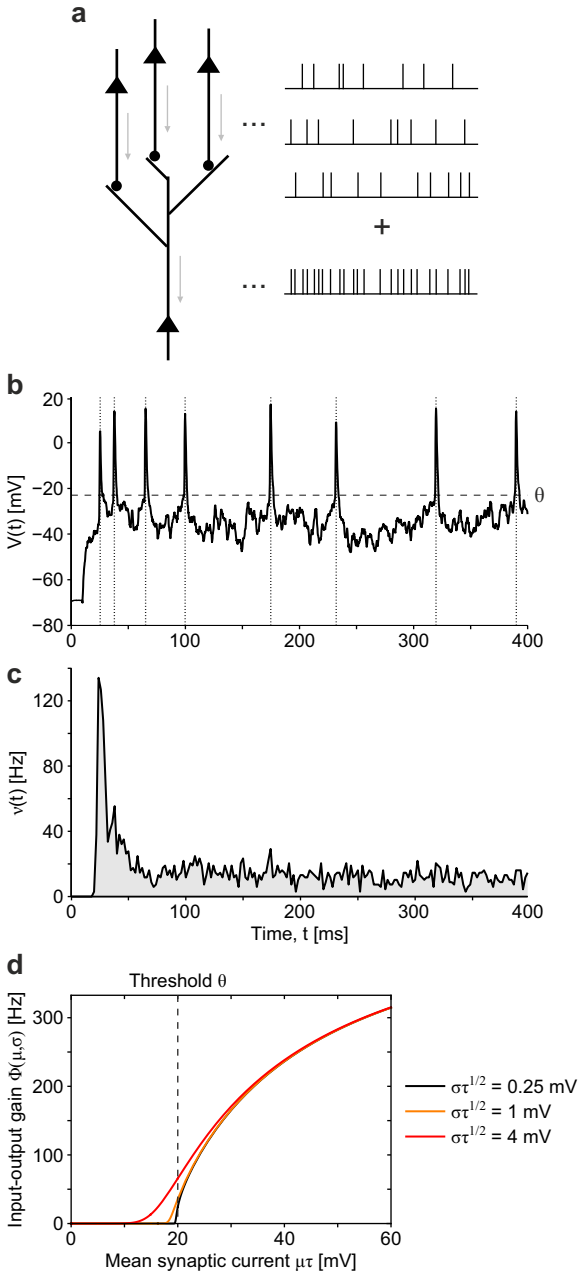


Figura 2: *(a) Schema di neurone e della composizione degli spike presinaptici in un unico treno Poissoniano. Sequenze di linee verticali, treni d'impulsi emessi da un neurone. (b) Simulazione di un neurone IF che riceve un treno di impulsi Possioniani in uno stato bilanciato tra eccitazione e inibizione. θ, soglia di emissione. (c) Frequenza $\nu(t)$ degli spike emessi da una rete di neuroni statisticamente equivalenti a quello mostrato in (b). (d) Funzione di guadagno $\Phi(\mu, \sigma)$ che da la frequenza di sparo di un neurone con corrente d'input di media μ e varianza σ^2. τ è la costante di rilassamento di $V(t)$ nel neurone IF. Pannello (a) adattato da [2].*

van Vreeswijk e Haim Sompolinsky [4], tale regime di emissione guidato principalmente dalle fluttuazioni delle corrente sinaptica può essere

prodotto in modo auto-consistente da una rete neuronale, e va sotto il nome di 'stato bilanciato' tra eccitazione e inibizione.

Il limite in cui le efficacie sinaptiche J sono molto più piccole della soglia di emissione θ, oltre ad avvicinarsi a quanto misurato sperimentalmente, permette di descrivere teoricamente il potenziale $V(t)$ come un processo stocastico di diffusione che segue l'equazione di Langevin

$$\frac{dV}{dt} = -\frac{V(t)}{\tau} + \mu(t) + \sigma(t)\,\xi(t)\,,$$

dove $\xi(t)$ è un rumore Guassiano senza memoria con media nulla e varianza unitaria. In condizioni stazionarie questo è il cosiddetto processo di Ornstein–Uhlenbeck, noto modello della velocità di una particella Browniana in un mezzo viscoso. Per semplicità consideriamo il solo caso in cui la trasmissione sinaptica è istantanea, cioè in cui l'arrivo di uno spike presinaptico induce una variazione di V su una scala di tempo molto breve rispetto a τ, e che i momenti della corrente non siano dipendenti dal potenziale di membrana stesso. Queste complicazioni che rendono la modellistica del neurone più aderente alla realtà biologica non cambiano qualitativamente il comportamento collettivo delle reti neuronali, ma in linea di principio possono, e spesso vengono, incluse come una generalizzazione del formalizzazione teorica qui riportata.

Nel descrivere la dinamica collettiva di una rete di neuroni siffatti, si può assumere che, sulla scala di interesse di una colonna corticale, in un dato strato, le proprietà statistiche delle correnti sinaptiche che ricevono siano approssimativamente uguali per tutti i neuroni. La corrente sinaptica in ogni cellula nervosa di una rete omogenea (eccitatoria o inibitoria) di una colonna corticale sarà allora una realizzazione diversa dello stesso processo stocastico. Questa è la cosiddetta approssimazione di 'campo media estesa' [3], ed è giustificata dal fatto che essendo il numero di contatti sinaptici K molto grande, lo scostamento dal suo numero medio sarà relativamente piccolo, come ci assicura la legge dei grandi numeri. I neuroni quindi sentiranno l'effetto dell'attività prodotta dalle altre $N-1$ cellule della rete attraverso la frequenza $\nu(t)$ data dal numero totale degli impulsi emessi al tempo t per unità di neurone (cioè diviso per N). La $\nu(t)$, di cui è illustrato un esempio in Figura 2c, è una variabile di stato collettiva e l'individualità dei neuroni a questo livello si perde. È da notare che anche se il singolo impulso emesso da un neurone presinaptico ha un effetto trascurabile nella dinamica di $V(t)$, due neuroni hanno comunque un'attività correlata, poiché le correnti (il 'campo') che ricevono hanno momenti che dipendono nello stesso modo dall'attività collettiva $\nu(t)$ (la 'media', da cui 'approssimazione di campo medio'). Questo non contraddice l'ipotesi di indipendenza invocata sopra, perché tale indipendenza si intende *condizionata* ai momenti (variabili nel tempo) dell'input sinaptico.

Nell'ambito della teoria dei processi stocastici citati, la dinamica di una popolazione neuronale può essere descritta in modo statisticamente completo seguendo l'evoluzione nel tempo della densità $p(v,t)$ di cellule che hanno potenziale di membrana in un piccolo intorno di v al tempo t [2]. A questo scopo è sufficiente ricorrere ad una equazione di continuità che descriva come varia nel tempo il numero di realizzazioni/neuroni $p(v,t)\,dv$ in un volumetto dv in funzione della divergenza del flusso $S(v,t)$ di realizzazioni che vi transitano. Nel caso unidimensionale l'equazione assume la forma $\partial_t p(v,t) = -\partial_v S(v,t)$. Per i processi stocastici di diffusione di cui stiamo parlando, questa equazione si esplicita nella seguente equazione di Fokker-Planck:

$$\partial_t p(v,t) = \partial_v[(\frac{v}{\tau} - \mu(t))\,p(v,t)] + \frac{1}{2}\sigma^2(t)\partial_v^2 p(v,t)\,.$$

Per quanto questo problema possa apparire come archetipico in fisica statistica, lo è solo in apparenza. In primo luogo le condizioni al contorno di questa equazione alle derivate parziali devono tenere conto del fatto che i neuroni che emettono un impulso al tempo t 'scompaiono' dal conteggio delle realizzazioni in $p(v,t)$ per poi 'ricomparire' nel potenziale di reset H dopo un periodo di inattività corrispondente al periodo refrattario assoluto. Questo porta ad avere un flusso di realizzazioni $S(v,t)$ diverso da zero anche in condizioni stazionarie, elemento questo tipico dei campi non conservativi. Il flusso alla soglia θ ci dice quindi la frazione di neuroni che emettono uno spike per unità di tempo, cioè $\nu(t) = S(\theta,t)$, che per media μ e varianza σ^2 costanti ci dice quanti spike per unità di tempo un generico neu-

rone della rete emetterà, fornendo una relazione di input-output tra la corrente d'ingresso ai neuroni e la loro attività, spesso indicata come funzione di guadagno $\Phi(\mu, \sigma)$, dal tipico aspetto sigmoidale. Alcuni esempi di questa funzione di guadagno sono illustrati in Figura 2d per diversi valori di σ, ovverosia della taglia delle fluttuazioni della corrente sinaptica. È interessante notare come per correnti quasi deterministiche (correnti con piccoli σ) i neuroni non sparano ($\Phi \simeq 0$) fino a quando la corrente media non supera il valore che permette al potenziale di membrana di attraversare la soglia di emissione ($\mu\tau > \theta$). All'aumentare di σ, e quindi delle fluttuazioni della corrente, le frequenza di output inizia a cresce anche per valori medi $\mu\tau$ del potenziale di membrana inferiori alla soglia: è questa la condizione in cui si manifesta lo stato bilanciato tra eccitazione e inibizione.

Altro elemento che rende l'equazione di Fokker-Planck della densità $p(v, t)$ di difficile soluzione è la sua nonlinearità. Poiché la media e varianza della corrente sinaptica dipendono dalla frequenza di firing $\nu(t)$, e questa a sua volta dipende dalla densità $p(v, t)$ che in un intorno di θ determina il flusso $S(\theta, t)$, μ e σ saranno indirettamente funzione di p. L'operatore di Fokker-Planck definito come $L\,p \equiv -\partial_v S$ sarà quindi esso stesso funzione della densità: $L = L(p)$. Nel corso degli anni in aggiunta alle integrazioni numeriche di questa equazione, si sono moltiplicati approcci teorici perturbativi, uno dei quali basato sullo sviluppo spettrale dell'operatore di Fokker-Planck L [6]. Per molti aspetti questo approccio ricorda il metodo di Hartree-Fock usato per stimare grandezze fisiche d'interesse in sistemi quantistici, con la differenza principale che le autofunzioni di L usate nello sviluppo, dipendono in questo caso esplicitamente dall'attività $\nu(t)$ della rete stessa. Il vantaggio di questa rappresentazione è quella di esprimere la dinamica di popolazione come la combinazione di 'modi' con una gerarchia di scale temporali data dall'inverso della parte reale degli autovalori dell'operatore L. Nello stato bilanciato il modo più lento è associato ad un autovalore reale che permette di derivare un espressione relativamente semplice della dinamica di $\nu(t)$:

$$\tau_\nu \frac{d\nu}{dt} = \Phi(\nu) - \nu\,.$$

L'equazione della frequenza di emissione qui riportata è un'equazione differenziale ordinaria di primo ordine la cui nonlinearità risiede nella funzione di guadagno Φ e nel fatto che la costante di tempo τ_ν varia con l'attività della rete. Nella sua forma, questa equazione delle frequenze ricorda da vicino quella presentata all'inizio degli anni '70 da Hugh R. Wilson e Jack D. Cowan come dinamica fenomenologica delle reti nervose eccitatorie e inibitorie, dando un contributo determinante negli anni a venire allo sviluppo delle neuroscienze computazionali; notiamo però che nelle equazioni di Wilson-Cowan la scala di tempo per l'evoluzione di ν è un parametro fenomenologico assegnato arbitrariamente.

Le soluzioni stazionarie dell'equazione precedente per la dinamica di ν corrispondono alla condizione di autoconsistenza $\nu_0 = \Phi(\nu_0)$: in approssimazione di campo medio lo stato stazionario identifica la condizione in cui la frequenza media di sparo del generico neurone uguaglia quella dei suoi neuroni presinaptici. Se il punto fisso ν_0 della dinamica di ν identificato dalla condizione di autoconsistenza è stabile, uno scostamento temporaneo da ν_0 (per esempio dovuto ad un piccolo stimolo) è seguito da un rilassamento verso ν_0, la cui componente lenta è governata dalla dinamica di campo medio sopra illustrata: ν_0 è quindi un *attrattore* della dinamica. In certe condizioni (essenzialmente, forte auto-eccitazione e conseguentemente una funzione di guadagno Φ fortemente nonlineare) l'equazione di autoconsistenza può avere più soluzioni (punti fissi) ν_0 (nel caso unidimensionale tre soluzioni, di cui due stabili ed una instabile nel mezzo). In questo caso, una perturbazione abbastanza forte a partire da un punto fisso stabile può spingere il sistema a 'scavalcare' il punto fisso instabile e rilassare nell'altro punto fisso stabile. Illustreremo nel seguito questo scenario in un contesto specifico.

Primitive computazionali più complesse: prendere decisioni

Prendere una *decisione* è, nell'accezione comune, sinonimo di compiere una scelta, e all'interno di questa definizione generale si può far rientrare gran parte del comportamento volontario (auto-

generato o innescato da stimoli esterni), dalla decisione di allontanare la mano da una fiamma a quella riguardante l'opportunità di un investimento finanziario. Ad un simile livello di generalità, la *percezione* indica qualunque processo fisico-fisiologico che converte segnali provenienti dal mondo esterno in specifici pattern di attività nervosa nel cervello.

La ricerca sulle basi neuronali dei processi di decisione copre un orizzonte molto ampio, e arriva a intersecare questioni psicologiche riguardanti l'attribuzione di valore alle opzioni disponibili, e perfino filosofiche riguardo al libero arbitrio. Qui ci limitiamo qui all'ambito ristretto della *decisione percettiva*, nella cui descrizione teorica molti dei concetti che abbiamo descritto giocano un ruolo importante.

In molte circostanze, la percezione avviene in modo quasi (o del tutto) inconsapevole, e semplici decisioni conseguenti sono affini a dei riflessi e non ingaggiano funzioni cognitive. A volte però l'ambiguità degli stimoli percettivi, e/o la criticità delle scelte conseguenti, richiedono invece l'elaborazione di processi di decisione.

Stiamo guidando nel traffico, di sera e sotto la pioggia; ad un tratto vediamo con la coda dell'occhio qualcosa, che potrebbe segnalare un imminente ostacolo e suggerire di frenare. Non vogliamo rischiare di investire qualcuno o qualcosa, ma non vogliamo rischiare inutilmente di sbandare o farci tamponare con una frenata di emergenza se si tratta di un falso allarme: dobbiamo decidere, e dobbiamo farlo sulla base della convinzione che riusciamo a farci sulla natura del potenziale ostacolo, e di un vincolo temporale stretto. La decisione percettiva è il processo attraverso il quale accumuliamo informazione (inclusa quella sensoriale e quella relativa al passaggio del tempo) per compiere una scelta.

La modellistica dei processi di decisione ha una lunga storia [5], e modelli astratti di questo processo sono stati proposti già negli anni '60. Parte della ricerca in questo ambito, di cui non ci occuperemo, riguarda ad esempio la formalizzazione dei principi di ottimalità delle decisioni in relazione al livello di incertezza sullo stato del mondo esterno (modelli Bayesiani).

In seguito ad una serie di importanti esperimenti negli anni '90, in cui si sono messi in luce alcuni meccanismi neurofisiologici della decisione percettiva, a partire dagli anni 2000 un ricco filone di ricerca teorica si è occupato della costruzione di modelli miranti a riprodurre, nella dinamica di popolazioni neuronali interagenti, alcune caratteristiche sia dell'attività neuronale osservata che delle misure psicofisiche effettuate nelle stesse condizioni.

I celebri esperimenti citati utilizzano un ingegnoso paradigma basato sui 'cinetogrammi a puntini' (traduzione italiana poco usata di 'random dots kinetograms'): nella versione più semplice, il soggetto guarda uno schermo sul quale sono proiettati molti puntini luminosi in movimento, parte verso sinistra e parte verso destra, e lo sperimentatore controlla la percentuale ('coerenza') di puntini che si muovono nella stessa direzione. Il compito consiste nel comunicare (attraverso un brusco cambiamento della direzione dello sguardo - 'saccade') la decisione percettiva del soggetto sulla direzione verso la quale si muove la maggioranza dei puntini (Figura 3a). È chiaro che per elevata coerenza il compito è facile, per il 50% di coerenza il soggetto può solo tirare a caso, e livelli intermedi di coerenza modulano la difficoltà del compito. In due versioni base del protocollo sperimentale, il soggetto è lasciato libero di prendere e comunicare la decisione appena si sente pronto a farlo, oppure lo sperimentatore fissa un tempo al quale la decisione deve comunque essere presa. Dal punto di vista psicofisico, tipicamente si misurano, per diversi livelli di coerenza, il tasso di errori ed il tempo di decisione (nella prima versione del compito), o il tasso di errori in funzione del tempo fissato per la decisione (nella seconda versione del compito) (Figura 3b).

Gli esperimenti identificarono nell'area 'intraparietale laterale' (LIP) popolazioni di neuroni la cui attivazione si correla in modo non ambiguo alla presa di decisione (nel senso che è possibile dissociare tale attività sia dalla percezione della direzione di movimento che dalla codifica dell'atto motorio attraverso il quale la decisione viene comunicata). Questa attività, mediata su molte ripetizioni del protocollo sperimentale, appariva crescere in modo più o meno costante dalla comparsa dello stimolo alla presa di decisione, e raggiungeva un valore fissato al momento della decisione (il che implica una crescita più ripida dell'attività per decisioni semplici e rapide, meno

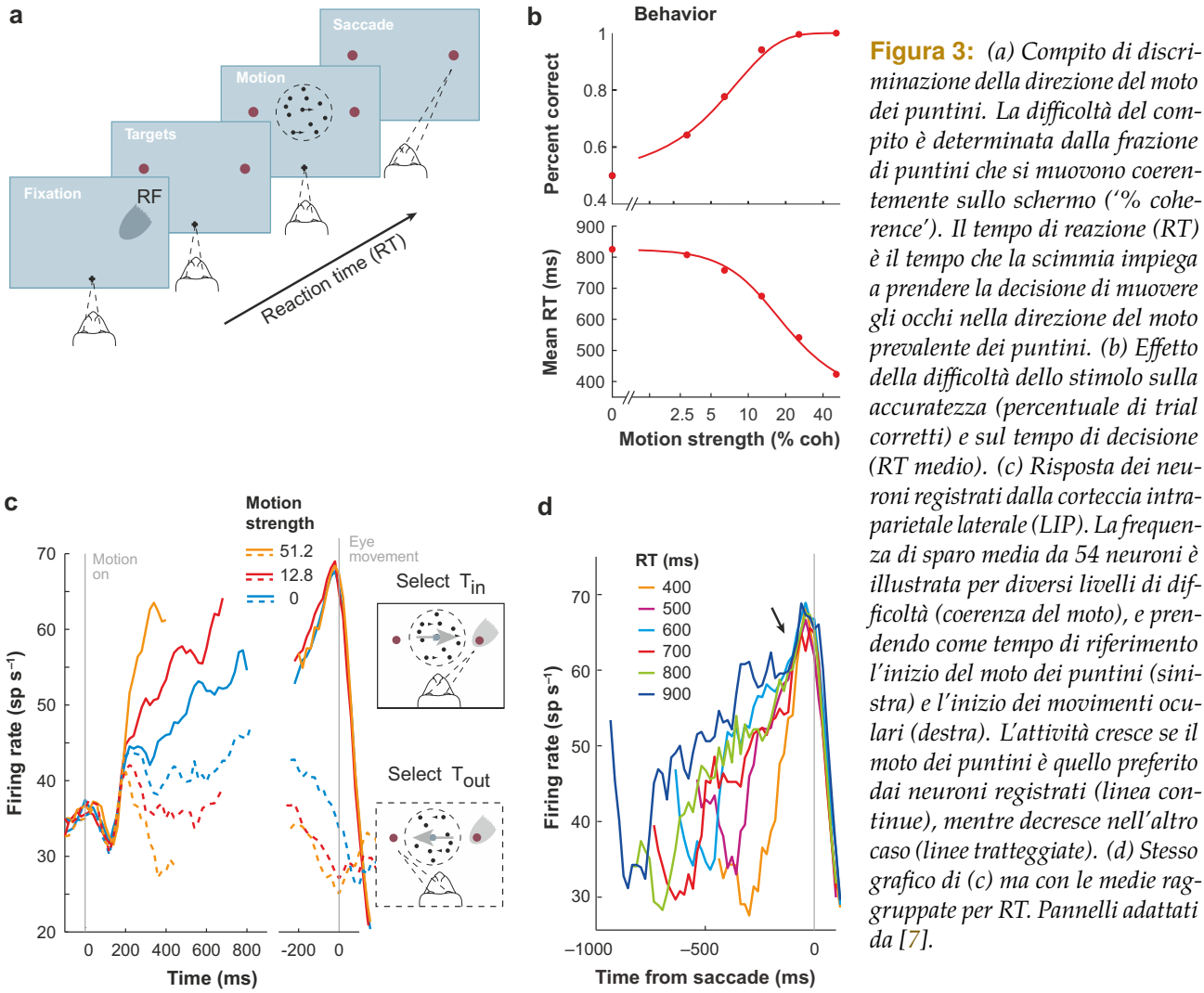


Figura 3: *(a) Compito di discriminazione della direzione del moto dei puntini. La difficoltà del compito è determinata dalla frazione di puntini che si muovono coerentemente sullo schermo ('% coherence'). Il tempo di reazione (RT) è il tempo che la scimmia impiega a prendere la decisione di muovere gli occhi nella direzione del moto prevalente dei puntini. (b) Effetto della difficoltà dello stimolo sulla accuratezza (percentuale di trial corretti) e sul tempo di decisione (RT medio). (c) Risposta dei neuroni registrati dalla corteccia intraparietale laterale (LIP). La frequenza di sparo media da 54 neuroni è illustrata per diversi livelli di difficoltà (coerenza del moto), e prendendo come tempo di riferimento l'inizio del moto dei puntini (sinistra) e l'inizio dei movimenti oculari (destra). L'attività cresce se il moto dei puntini è quello preferito dai neuroni registrati (linea continue), mentre decresce nell'altro caso (linee tratteggiate). (d) Stesso grafico di (c) ma con le medie raggruppate per RT. Pannelli adattati da [7].*

ripida per decisioni difficile e lente) (Figura 3c).

Coerentemente con l'intuizione e con modelli astratti precedenti, il quadro suggeriva un processo di accumulo di informazione (tanto più lento e lungo quanto più lo stimolo è ambiguo), fino a che il raggiungimento di una soglia rende possibile la decisione (Figura 3d).

Nel modello proposto da X.-J. Wang nel 2002 [8], ciascuna delle due decisioni disponibili ('destra' o 'sinistra') è rappresentata dall'attrattore dinamico di una popolazione eccitatoria ricorrente (nel senso spiegato alla fine della sezione precedente), e il sistema è composto da due popolazioni ricorrenti eccitatorie i cui attrattori codificano le due scelte 'selettive', una popolazione eccitatoria non direttamente coinvolta nella rappresentazione delle due scelte e da una popolazione di neuroni inibitori (Figura 4a). Le popolazioni eccitatorie forniscono input alla popolazione inibitoria, che a sua volta fornisce ad esse un feedback inibitorio. Una configurazione di questo tipo realizza un meccanismo competitivo 'winner-take-all': se l'attività di una delle due popolazioni selettive supera quella dell'altra in misura superiore ad una soglia, la reazione inibitoria amplifica ulteriormente questa differenza, finché si raggiunge uno stato stabile in cui la popolazione eccitatoria 'vincente' e quella 'perdente' si attestano rispettivamente su uno stato di alta e bassa attività. Gli attrattori della dinamica complessiva di questo sistema sono dunque i due stati corrispondenti alla 'vittoria' dell'una o dell'altra popolazione.

Nella descrizione del processo di decisione, il sistema parte da una condizione simmetrica in cui entrambe le popolazioni eccitatorie sono in uno stato di bassa attività (attività spontanea). L'accensione dell'input visivo corrisponde all'attivazione di due input (rumorosi) a ciascuna delle popolazioni eccitatorie corrispondenti la cui

intensità media codifica il livello di coerenza per le due direzioni di movimento dei puntini. Nel caso massimamente ambiguo (50% di coerenza) la media dei due input sarà uguale, l'attività delle due popolazioni eccitatorie crescerà in media in modo uguale, guidata dalla risposta nonlineare di auto-eccitazione. La crescita sarà temperata, ma in modo uguale, dalla reazione inibitoria. Data la componente stocastica dell'input, l'attività crescente corrisponderà a fluttuazioni di dimensione crescente tra le attività delle due reti eccitatorie, finché una fluttuazione sarà sufficiente a determinare una reazione inibitoria in grado di rompere la simmetria e innescare in modo irreversibile la dinamica competitiva che porterà una delle due popolazioni alla vittoria. In questo caso, il ruolo delle fluttuazioni casuali nel determinare la decisione corrisponde ovviamente al fatto che per coerenza al 50% il soggetto può solo tirare a caso.

Diversi livelli di coerenza corrispondono a diversi sbilanciamenti negli input ricevuti dalle due popolazioni eccitatorie, e questo fornisce una componente deterministica che rompe dall'inizio la simmetria tra le due e produrrà decisioni più rapide e più affidabili al crescere dell'asimmetria.

Prima di illustrare brevemente il tipo più semplice di modello proposto per descrivere la dinamica neuronale della decisione percettiva, sottolineiamo che la crescita graduale dell'attività neuronale media osservata, a cui abbiamo accennato sopra, si sviluppa (a seconda della difficoltà del compito) su scale di tempo che arrivano facilmente al secondo ed oltre. Ottenere una dinamica così lenta in reti i cui neuroni hanno tempi caratteristici molto più brevi è difficile, e questo pone una sfida ai modelli teorici. Nel modello proposto in [8], e in diversi lavori seguenti, si enfatizza il possibile ruolo della componente più lenta della trasmissione sinaptica eccitatoria (che corrisponde all'attivazione dei cosiddetti recettori NMDA), la cui scala di tempo si stima in decine di ms. Accenneremo all'esistenza di opzioni diverse, riguardo sia al meccanismo dinamico di decisione che all'interpretazione degli andamenti osservati negli esperimenti.

Nello spirito di una descrizione di campo medio, identifichiamo lo stato del sistema ad un certo istante con i corrispondenti valori delle frequenze medie di emissione di spike delle popolazioni coinvolte. Per le due popolazioni eccitatorie A e B che codificano la decisione sulla direzione di movimento possiamo scrivere una dinamica di campo medio (Figura 4b):

$$\begin{aligned} \tau_E \dot{\nu}_A &= -\nu_A + \Phi_E(J_{EE}\nu_A - J_{EI}\nu_I + I_A) \\ \tau_E \dot{\nu}_B &= -\nu_B + \Phi_E(J_{EE}\nu_B - J_{EI}\nu_I + I_B) \end{aligned}$$

A queste equazioni si dovrebbero aggiungere un'equazione analoga per l'attività media ν_I della popolazione inibitoria, caratterizzata da una scala di tempo τ_I. Per focalizzarci sulla dinamica delle due popolazioni eccitatorie, ν_A e ν_B (il che ci permette di utilizzare gli strumenti di analisi dei sistemi dinamici sul 'piano di fase') assumiamo che $\tau_I \ll \tau_E$; in altre parole ipotizziamo che la dinamica della popolazione inibitoria sia così veloce rispetto a quella eccitatoria che sulla scala di tempo τ_E essa raggiunga istantaneamente la condizione di equilibro $\nu_I = \Phi_I(\nu_I, \nu_A, \nu_B)$ rispetto ai valori correnti di ν_A e ν_B. Questo ci permette di sostituire ν_I con $\Phi_I(\nu_I, \nu_A, \nu_B)$ nelle precedenti equazioni. Questa drastica semplificazione è parzialmente giustificata dal fatto che tipicamente i neuroni inibitori generano spike a frequenza più alta di quelli eccitatori, e che gli input sinaptici inibitori sono mediati da recettori veloci.

Spesso si adotta un'altra approssimazione, non necessaria alla riduzione bidimensionale del modello, che semplifica le equazioni e consiste nell'assumere che la funzione di trasferimento Φ_I dei neuroni inibitori sia lineare ($\nu_I = \alpha\,(\nu_A + \nu_B)$), in modo tale che le equazioni dinamiche del modello si riducono a:

$$\begin{aligned} \tau_E \dot{\nu}_A &= -\nu_A + \Phi_E\left(\tilde{J}_{EE}\nu_A - J_{EI}\alpha\nu_B + I_A\right) \\ \tau_E \dot{\nu}_B &= -\nu_B + \Phi_E\left(\tilde{J}_{EE}\nu_B - J_{EI}\alpha\nu_A + I_B\right) \end{aligned}$$

con $\tilde{J}_{EE} \equiv J_{EE} - \alpha J_{EI}$. Questo sistema di equazioni può essere analizzato con strumenti standard dello studio dei sistemi dinamici bidimensionali, che permettono di studiare molte caratteristiche della dinamica senza dover ricavare esplicitamente le soluzioni delle equazioni (un ottimo riferimento è il testo di S. Strogatz [9]). In breve: nel piano generato da ν_A e ν_B si tracciano innanzitutto le due cosiddette *nullcline*, luoghi dei punti in cui $\dot{\nu}_A = 0$ e $\dot{\nu}_B = 0$, rispettivamente.

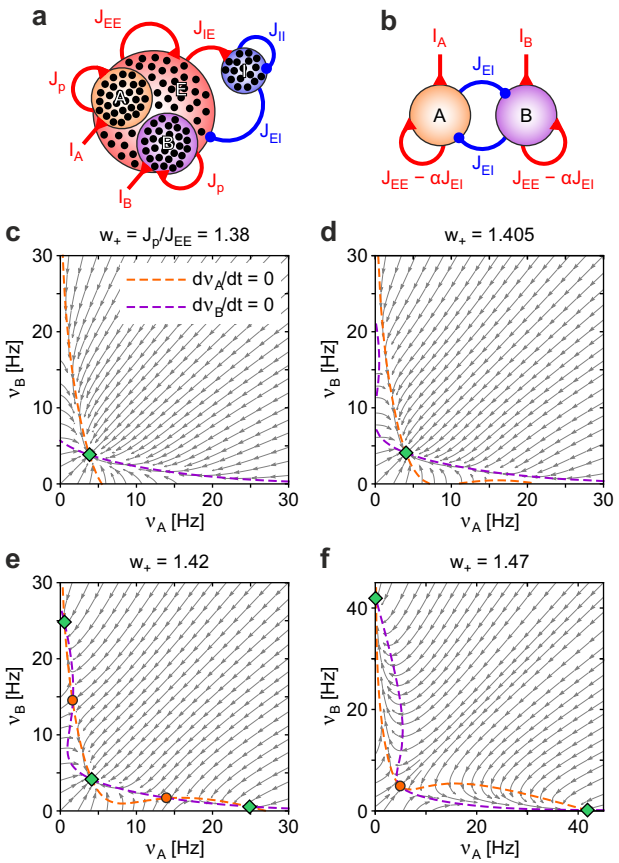


Figura 4: *(a) Rete di neuroni IF (puntini neri) modello del modulo decisionale in LIP. I neuroni eccitatori sono divisi in tre gruppi: due codificano la scelta A o B, rispettivamente, e uno non è coinvolto attivamente nel compito (E). I neuroni in A e B ricevono un input I_A e I_B esterno associato alla frazione di punti che si muove nelle rispettive direzioni. Le efficacie sinaptiche $J_{\alpha\beta}$ indicano la forza di accoppiamento tra neuroni delle popolazioni α e β. (b) Riduzione a due popolazioni del modello in approssimazione di campo medio (vedi testo). (c-f) Piani delle fasi (ν_A, ν_B) delle frequenze di sparo della rete in (b) per diversi livelli di auto-accoppiamento sinaptico $w_+ = J_p/J_{EE}$. Linee tratteggiate, nullcline della dinamica di ν_A e ν_B. Frecce, linea di flusso del campo di velocità. Cerchi rossi, punti fissi instabili. Rombi verdi, fuochi stabili.*

Le nullcline forniscono una sorta di sistema di riferimento sul piano di fase, che lo partiziona in regioni all'interno delle quali le componenti del vettore $(\dot{\nu}_A, \dot{\nu}_B)$ hanno segno definito; in questo modo si ricava un quadro complessivo delle direzioni del flusso vettoriale rispetto alle nullcline. I punti di intersezione delle nullcline, se esistono, sono punti fissi della dinamica, in quanto soddisfano la condizione $\dot{\nu}_A = \dot{\nu}_B = 0$. Un punto fisso può essere stabile o instabile, tale cioè che, partendo da una condizione iniziale ad esso vicina (analisi lineare), la dinamica riporta il sistema al punto fisso o lo allontana, rispettivamente. È possibile che un punto fisso risulti stabile rispetto ad una direzione di allontanamento e stabile rispetto ad un'altra (queste direzioni sono identificate dagli autovettori della matrice che linearizza la dinamica): si tratta in questo caso di un 'punto di sella'. L'analisi di stabilità descrive la dinamica linearizzata del sistema in prossimità del punto fisso; se questo risulta instabile, la dinamica a lungo termine del sistema va analizzata con altri strumenti. È comunque possibile ricavare dall'analisi lineare alcune informazioni 'anticipatorie' del destino a lungo termine del sistema, ad esempio nella distinzione tra una fuga rettilinea ed una spirale divergente.

Uno dei vantaggi della riduzione bidimensionale di un sistema dinamico è che in due dimensioni il repertorio di regimi dinamici a lungo termine si riduce a due casi (due tipologie di attrattore): punti fissi o 'cicli limite' (orbite chiuse attrattive). Al di là dell'analisi lineare, ogni attrattore (punto fisso o ciclo limite) è associato ad un 'bacino di attrazione', cioè alla regione del piano di fase tale che per una condizione iniziale in esso contenuta la dinamica porta asintoticamente il sistema nell'attrattore.

Le equazioni che abbiamo scritto per il modello di decisione dipendono da diversi parametri, che a loro volta influenzano la forma delle nullcline e quindi la posizione e stabilità dei punti fissi, e la natura degli attrattori della dinamica.

Due grandezze giocano un ruolo chiave in questo scenario: l'entità dell'auto-eccitazione delle due popolazioni selettive, ed il livello di coerenza dello stimolo. Lo studio sistematico del modo in cui la dinamica del sistema cambia qualitativamente al variare di uno o più parametri è la *teoria delle biforcazioni*; senza pretesa di completezza, illustriamo le conseguenze principali di questa analisi. L'auto-eccitazione (quantificata dal potenziamento sinaptico relativo w_+ in Figura 4) determina la nonlinearità della risposta allo stimolo e, secondo quanto discusso alla fine della sezione precedente, il repertorio e la natura dei punti fissi della dinamica di campo medio, a parità di stimolo: in assenza di stimolo, per bassa auto-eccitazione l'unico punto fisso è un attrattore a bassa attività (Figura 4c). Al crescere

di w_+ ad un certo punto (di 'biforcazione pitchfork') compaiono, insieme al precedente punto fisso che rimane stabile, due altri punti fissi asimmetrici, caratterizzati da alta attività di una popolazione selettiva e bassa attività dell'altra (Figura 4e). Un ulteriore aumento di w_+ arriva a destabilizzare il punto fisso di bassa attività, lasciando come unici punti fissi stabili i due attrattori asimmetrici (Figura 4f). A seconda del valore di w_+, si possono avere diversi comportamenti dinamici al variare dello stimolo, esemplificati in Figura 5 nella quale, per due valori di w_+, si descrive la rappresentazione del sistema nel piano di fase prima, durante e dopo la stimolazione. Per il valore maggiore di w_+ (Figura 5a), il sistema possiede, come abbiamo visto sopra, un punto fisso simmetrico stabile di bassa attività, che costituisce la condizione pre-stimolazione, e due punti fissi stabili asimmetrici. Durante la stimolazione, di coerenza tale da favorire uno degli attrattori asimmmetrici, il paesaggio di fase viene temporaneamente deformato insieme ad i suoi punti fissi, ed il sistema viene spinto verso una versione 'potenziata' dell'attrattore asimmetrico favorito dallo stimolo. Dopo la stimolazione il paesaggio di fase torna alla condizione pre-stimolo, ma il sistema ora rimane 'intrappolato' nell'attrattore asimmetrico. Per un basso valore di w_+ (Figura 5b), il paesaggio pre-stimolo espone un solo punto fisso stabile a bassa attività e, sebbene la stimolazione spinga temporaneamente il sistema verso uno stato asimmetrico, nella fase post-stimolo il sistema rilassa verso la stessa condizione pre-stimolo.

I modelli di decisione percettiva del tipo descritto in [8] si basano sul fatto che l'arrivo dello stimolo (il display dei puntini in movimento) modifica il paesaggio del sistema nel piano di fase in modo da destabilizzare lo stato simmetrico di bassa attività, trasformandolo in una sella. In assenza di rumore e per coerenza nulla il sistema 'scivolerebbe' sul crinale della soglia, continuando a rispettare la simmetria, ma anche una modesta fluttuazione romperebbe la simmetria e dal momento in cui il sistema lascia il crinale della soglia la sua dinamica è determinata dalla pendenza del paesaggio che lo accompagna verso uno dei due attrattori asimmetrici. Per coerenza non nulla, il sistema viene subito allontanato dal crinale della sella, e la traiettoria verso l'attrattore asimmetrico è dominata dalla dinamica competitiva descritta e guidata in modo essenzialmente deterministico dalla forma del paesaggio; l'eventuale rumore presente ha un effetto marginale.

Notiamo di passaggio che è possibile stabilire una corrispondenza formale tra i modelli bidimensionali appena descritti e la categoria di modelli che sopra abbiamo definito 'astratti', unidimensionali, essenzialmente basati su un meccanismo di integrazione (con rumore) del segnale di input in cui la decisione corrisponde all'attraversamento di una soglia da parte dell'integratore. In questa corrispondenza, l'innesco del meccanismo winner-take-all del modello bidimensionale corrisponde al raggiungimento della soglia.

Tornando ai modelli bidimensionali, il tempo di decisione, cioè il tempo tra la reazione allo stimolo al raggiungimento dell'attrattore, è influenzato dai tempi caratteristici della trasmissione sinaptica, e come abbiamo accennato è stato originariamente proposto che la scala di tempo, comparativamente lunga, del recettore NMDA, fosse l'elemento chiave per dar conto dei tempi di decisione osservati. È però possibile immaginare, per gli stessi modelli descritti, un diverso meccanismo dinamico, che ha il pregio di assicurare in modo naturale una notevole flessibilità nei tempi di decisione su range molto ampi, e fornisce anche un supporto teorico ad alcune recenti analisi dell'attività neurale durante la decisione. L'idea è semplice: uno stimolo non abbastanza forte da destabilizzare il punto fisso di bassa attività ne diminuirà comunque la stabilità (cioè modificherà comunque il paesaggio, diminuendo la profondità della 'valle' di bassa attività). In assenza di rumore quindi il sistema rimarrebbe ancorato a questo punto fisso, ma il rumore può favorirne la fuga nella valle corrispondente ad uno dei due attrattori asimmetrici. In fisica questo tipo di situazione corrisponde ad un problema noto (*noise-driven transitions*), ed un risultato fondamentale è che la probabilità per unità di tempo di fuga da una valle dipende in modo esponenziale dal rapporto tra l'altezza della barriera che dev'essere superata e la varianza del rumore (equazione di Arrhenius). Nel nostro caso, l'altezza della barriera è la quantità che viene modulata dall'intensità dello stimolo, e si capisce quindi come questa dipendenza esponenziale

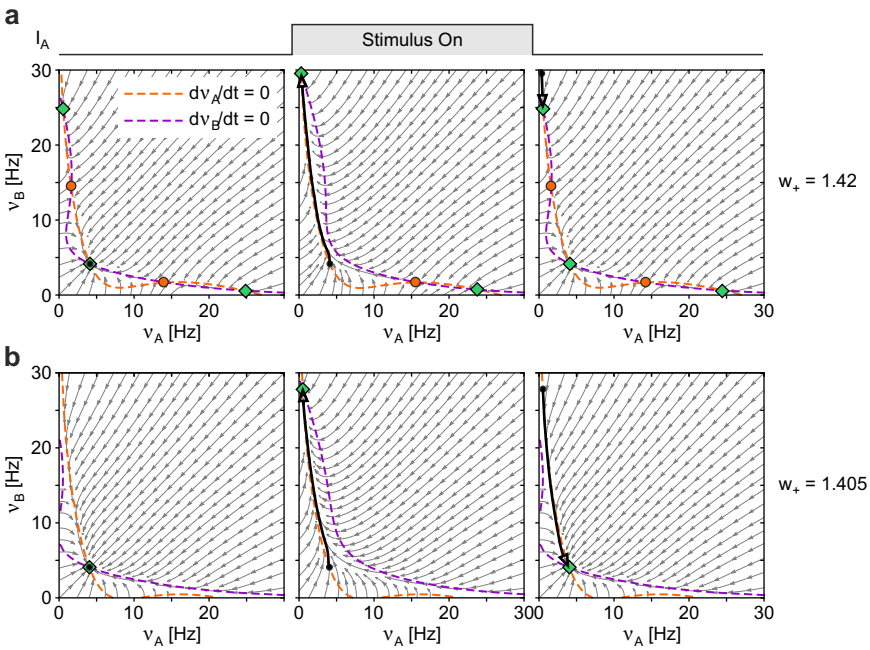


Figura 5: *Piani delle fasi (ν_A, ν_B) della rete in Figura 4b con l'aggiunta della traiettoria da essa seguita (curve con freccia nere). Le tre fasi del compito illustrate sono: i) il periodo che precede l'inizio del compito; ii) il periodo in cui appaiono i puntini in movimento nella direzione A portando ad un incremento della sola I_A; iii) la fase finale del compito in cui la stimolazione scompare. Le reti illustrate sono: (a) una con forte auto-accopiamento w_+ (tre attrattori stabili), e (b) l'altra con le sottopopolazioni A e B debolmente accoppiate (il solo attrattore a bassa frequenza ν è presente). Simboli non descritti, come in Figura 4.*

consenta di modulare il tempo medio di fuga (il tempo di decisione nel nostro caso) su un range molto ampio anche con modeste variazioni dell'intensità dello stimolo. Non è facile estendere la trattazione analitica classica del problema unidimensionale al caso in due dimensioni (sebbene sia possibile una riduzione approssimata ad un caso unidimensionale in prossimità della biforcazione), e l'opzione descritta è stata studiata in [10] essenzialmente attraverso simulazioni numeriche, caratterizzando la dipendenza della distribuzione dei tempi di decisione dall'intensità dello stimolo e da quella del rumore.

In questo scenario, dunque, il rumore gioca un ruolo costruttivo essenziale. Se da un lato l'inclusione del rumore è del tutto naturale, per le motivazioni discusse sopra, resta da vedere se i dati sperimentali offrono un chiaro supporto all'uno o l'altro scenario.

La questione è articolata e complessa, e non possiamo discuterla qui in dettaglio; ci focalizziamo su un solo aspetto, che riveste un interesse generale al di là del processo di decisione oggetto dei modelli.

Abbiamo ricordato che, dalle registrazioni elettrofisiologiche effettuate durante il processo di decisione, l'attività dei neuroni coinvolti cresce lungo una 'rampa' fino a raggiungere un valore che sembra esprimere una soglia, e che la pendenza della rampa, non la soglia, sembra modulata dalla difficoltà del compito (dalla coerenza). Bisogna però sottolineare che queste 'rampe' vengono chiaramente osservate quando si effettua la media dell'attività neuronale su molte ripetizioni del protocollo sperimentale ('trial'); riconoscerle nell'attività durante una singola ripetizione è molto più difficile. D'altra parte, il meccanismo guidato dal rumore appena descritto non predice certo 'rampe' nell'attività neuronale: questa rimarrebbe bassa e stazionaria finché il sistema non viene spinto da una fluttuazione oltre la barriera, e dopo la fuga aumenterebbe rapidamente fino ad attestarsi intorno al valore elevato corrispondente ad uno dei due attrattori asimmetrici; invece di una rampa, quasi un 'gradino'. Ma essendo ora il processo guidato dal rumore, a parità di input l'istante di tempo in cui il sistema supera la barriera varierà molto (con una distribuzione approssimativamente esponenziale). La 'rampa' osservata negli esperimenti potrebbe essere quindi una conseguenza (in un certo senso, un artefatto) di una media effettuata su attività neuronali che, nel singolo trial, esprimono un andamento 'a gradino'; in [10] viene esemplificata la plausibilità di questo meccanismo (Figura 6). Un esempio di questi due tipi di attività osservabile a livello di singolo trial è stata illustrata non solo in corteccia parietale [11] ma anche nelle cortecce motorie [12]. Decidere tra le due opzioni sulla base dei dati può apparire facile ma non lo è, e il dibattito è ancora acceso (si veda ad esempio [13]).

Notiamo che, indipendentemente dalla formulazione dinamica dei modelli, l'analisi comparati-

va delle due ipotesi si può formulare puramente in termini di analisi statistica del segnale, in cui si quantifica la verosimiglianza dei due modelli statistici (la probabilità condizionata di ottenere il dato osservato, dato il modello statistico descrittivo adottato).

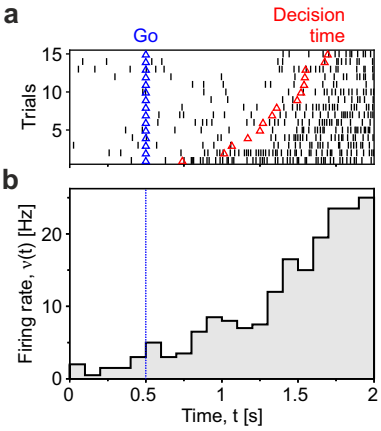


Figura 6: *Dinamica di singolo neurone nella rete in Figura 4a in regime di decisione guidata dal rumore. (a) Spike emessi dallo stesso neurone della rete simulata nei diversi trial. Ogni trattino verticale è uno spike. Simboli verdi e rossi rappresentano l'inizio della stimolazione (blue) e il tempo in cui viene presa la decisione (rosso). (b) Istogramma normalizzato (cioè la frequenza di emissione $\nu(t)$) degli spike emessi ed illustrati in (a). Adattata da [10].*

Si diceva della generalità di alcune implicazioni dei meccanismi dinamici discussi; un esempio stimolante è fornito da una categoria di fenomeni percettivi curiosi e per certi versi paradossali, esemplificati dalla 'rivalità binoculare' e dalla percezione degli stimoli ambigui. Di questi ultimi conosciamo quasi tutti degli esempi (la 'giovane-vecchia, il cubo di Necker ecc.). La rivalità binoculare è il fenomeno percettivo per cui se due immagini vengono mostrate, simultaneamente e separatamente, ai due occhi, la nostra percezione visiva cosciente alterna (e in modo apparentemente casuale) tra la piena 'visione' dell'una o dell'altra immagine (invece di percepire, come pure verrebbe naturale supporre, un mescolamento delle due immagini). Si osserva che gli intervalli di tempo di 'dominanza' di ciascuna delle due percezioni coscienti hanno una distribuzione statistica che mantiene invariate alcune caratteristiche anche per condizioni sperimentali molto diverse, a suggerire che un meccanismo comune potrebbe esserne alla base [14]. Si noti che la variabilità statistica di questi processi si manifesta a fronte di uno stimolo costantemente presente e invariato, il che quasi irresistibilmente spinge a pensare ad un meccanismo stocastico interno.

Questi fenomeni percettivi sono studiati da decenni in psicofisica, e diversi modelli sono stati formulati per tentarne una spiegazione; senza tentarne una rassegna, per la nostra discussione è interessante osservare che è possibile spiegare diverse caratteristiche statistiche dell'alternanza percettiva attraverso modelli teorici che di fatto estendono il modello di decisione basato sulle fluttuazioni che abbiamo appena discusso [15],[14]; in effetti, l'alternanza percettiva può essere vista come una 'decisione involontaria'.

Sottolineiamo infine alcune assunzioni/limitazioni importanti, sia degli esperimenti che dei modelli di decisione percettiva: *i)* la scelta è obbligatoria, il repertorio di scelte è fissato e le scelte disponibili sono due (*two-alternative forced choice task*; questa limitazione può apparire banale ma non lo è); *ii)* l'input alle due popolazioni che codificano la decisione è completamente segregato in partenza rispetto alla direzione di movimento; *iii)* si assume che la rappresentazione neuronale delle decisioni disponibili sia pre-formata e il modello non si occupa di come essa si possa generare dall'esperienza (fedeli all'impegno di non trattare i processi di apprendimento, non discuteremo questo punto).

Conclusioni

Lo sviluppo di modelli fisico-matematici dell'attività nervosa ha una lunga storia (il fondamentale modello di Hodgkin-Huxley della generazione dello spike è del 1952), ed il ruolo dei fisici e dei matematici in questo ambito ha costruito da subito un germe di interdisciplinarietà nello studio del cervello. D'altra parte, nell'ambito delle (allora giovani) discipline psicologiche, da subito si avvertì l'utilità metodologica di astrarre uno schema concettuale dei rapporti funzionali tra diverse componenti del sistema nervoso, disancorato dal dettaglio biologico (è il caso, per esempio, del 'sistema nervoso concettuale', a cui si riferirono Hebb, Skinner ed altri, e forse lo

schema delineato da Freud nel 'Progetto di una psicologia' ne è un antesignano).

I progressi degli ultimi decenni, però, stanno mutando qualitativamente il quadro. Da una situazione in cui il flusso culturale tra biologia e 'scienze dure', pur fecondo, manteneva i due ambiti nettamente distinti, con le seconde 'a servizio' della prima per sistematizzare attraverso modelli quantitativi la quantità di dati empirici raccolti, si stanno ora consolidando nuovi profili scientifici, offerte formative e contesti di ricerca, grazie ad una osmosi che sta creando una 'scienza del cervello' multi-disciplinare.

Una conseguenza di questa trasformazione è il fatto che lo studio del cervello pone nuovi problemi alle scienze fisiche, stimolando lo sviluppo di nuovi strumenti teorici e computazionali. Un aspetto di questo, che speriamo emerga almeno in parte da quanto scritto, è la natura multi-scala dello studio del cervello.

La fisica si è sempre avvantaggiata della possibilità di separare le scale nella descrizione di un fenomeno: la dinamica dei fluidi necessaria a progettare la chiglia di una barca non tiene conto della struttura molecolare dell'acqua, se non nella determinazione di alcuni parametri che entrano nella descrizione macroscopica del problema. Analogamente, è spesso importante poter separare le scale di tempo in modo tale che, in un sistema i cui gradi di libertà evolvono su scale di tempo molto diverse, si possa descrivere la dinamica dei gradi di libertà veloci nel background 'statico' di quelli lenti. I problemi fisici in cui la separazione di scale non è possibile sono notoriamente difficili, e la meccanica statistica dei fenomeni critici ne è un esempio.

Lo studio teorico del cervello aggiunge a questo quadro diversi livelli di complessità: diverse scale spaziali e temporali sono accoppiate in modo difficilmente separabile, ma non costituiscono un continuo (si pensi alla modularità spaziale di cui abbiamo discusso), e la molteplicità di scale non è realisticamente trattabile con gli strumenti della meccanica statistica dei fenomeni critici, adatti a studiare fluttuazioni multi-scala in sistemi omogenei. Il sistema in generale non è assimilabile ad un sistema fisico, per quanto complesso, all'equilibrio, e la fisica dei sistemi a molti corpi fuori dall'equilibrio è molto più difficile da trattare (un fisico trova facilmente libri e articoli di 'meccanica statistica di non-equilibrio', ma spesso il titolo apparentemente generale corrisponde ad una scelta idiosincratica di metodi adatti a situazioni specifiche - qualcuno ha scritto che definire la 'meccanica statistica di non-equilibrio' è un po' come definire la 'zoologia dei non-elefanti'). In relazione all'osservazione precedente, il cervello è chiaramente un 'sistema aperto', le cui interazioni col mondo esterno sono molto più difficili da schematizzare teoricamente di un 'bagno termico' o di un campo magnetico variabile.

Queste osservazioni suggeriscono che, dal punto di vista dello studio teorico, il cervello si colloca in una 'terra di mezzo'. Con ammirevole preveggenza, Warren Weaver nel 1948 [16] la identificò riflettendo sulle sfide inedite che i sistemi biologici complessi avrebbero posto alla fisica, e distingueva problemi 'semplici' (non vuol dire facili da risolvere, ma trattabili con astrazioni che semplificano la logica dei modelli, come nel calcolo deterministico delle orbite dei pianeti), e problemi di 'complessità disorganizzata' (per esempio quelli in cui le fluttuazioni casuali sono importanti ma si possono assumere omogenee nello spazio e nel tempo, e si riesce a descrivere il comportamento del sistema attraverso l'evoluzione delle distribuzioni di probabilità di poche osservabili). 'In mezzo', appunto, stanno i problemi di 'complessità organizzata'. Nella classificazione di Weaver, l'approccio teorico al cervello ricade senz'altro nell'ultima categoria.

Per i motivi che abbiamo sopra riassunto, la costruzione di modelli realistici di attività del cervello è difficile, e la loro soluzione analitica supera spesso le nostre capacità. Questo fatto motiva (come, di nuovo, lo stesso Weaver aveva predetto) un ruolo centrale delle simulazioni al calcolatore dei modelli. Nello studio teorico del cervello gli approcci computazionali hanno assunto un ruolo crescente (anche grazie, ovviamente, alle crescenti potenze di calcolo disponibili), ed il loro ruolo euristico è oggi spesso assimilato a veri e propri 'esperimenti *in silico*'; ne è testimonianza il fatto che gli sviluppi computazionali sono centrali in diverse iniziative progettuali multi-nazionali; in particolare nello Human Brain Project, iniziativa europea decennale a cui l'Italia dà un notevole contributo, che dopo la sua conclusione nel 2023 si evolverà in una in-

frastruttura europea (EBRAINS) che metterà a servizio della comunità scientifica il ricco repertorio di strumenti computazionali sviluppati nel progetto [17]. È auspicabile, e probabile, che questi sviluppi accelereranno in modo importante l'evoluzione della comunità multi-disciplinare impegnata nella affascinante frontiera costituita dallo studio del cervello.

[1] V. B. Mountcastle: *The columnar organization of the neocortex*, Brain, 120 (1997) 101.

[2] P. Del Giudice and M. Mattia: *Stochastic population dynamics of spiking neurons* In E. Korutcheva and R. Cuerno, editors, Advances in Condensed Matter and Statistical Physics, chapter 10, pages 125–153. Nova Science Publishers, New York, (2004).

[3] D. J. Amit and N. Brunel. *Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex.* Cereb. Cortex, 7 (1997) 237.

[4] C. van Vreeswijk and H. Sompolinsky. *Chaos in neuronal networks with balanced excitatory and inhibitory activity.* Science, 274 (1996) 1724.

[5] R. Ratcliff, P. L. Smith, S. D. Brown, and G. McKoon: *Diffusion decision model: Current issues and history* Trends in cognitive sciences, 20 (2016) 260.

[6] M. Mattia and P. Del Giudice. *Population dynamics of interacting spiking neurons*. Phys. Rev. E, 66 (2002) 051917.

[7] J. I. Gold and M. N. Shadlen: *The neural basis of decision making*. Annu. Rev. Neurosci. 30 (2007) 74.

[8] X.-J. Wang. *Probabilistic decision making by slow reverberation in cortical circuits.* Neuron, 36 (2002) 955.

[9] S. H. Strogatz: *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering*. Westview Press, (2000).

[10] D. Martí, G. Deco, M. Mattia, G. Gigante, and P. Del Giudice. *A fluctuation-driven mechanism for slow decision processes in reverberant networks.* PloS one, 3 (2008) e2534.

[11] T. A. Engel, N. A. Steinmetz, M. A. Gieselmann, A. Thiele, T. Moore, and K. Boahen. *Selective modulation of cortical state during spatial attention.* Science, 354 (2016) 1140.

[12] M. Mattia, P. Pani, G. Mirabella, S. Costa, P. Del Giudice, and S. Ferraina. *Heterogeneous attractor cell assemblies for motor planning in premotor cortex.* J. Neurosci., 33 (2013) 11155.

[13] D. M. Zoltowski, K. W. Latimer, J. L. Yates, A. C. Huk, and J. W. Pillow. *Discrete stepping and nonlinear ramping dynamics underlie spiking responses of lip neurons during decision-making* Neuron, 102 (2019) 1249.

[14] R. Cao, A. Pastukhov, M. Mattia, and J. Braun. *Collective activity of many bistable assemblies reproduces characteristic dynamics of multistable perception.* Journal of Neuroscience 36 (2016) 6957.

[15] G. Gigante, M. Mattia, J. Braun, and P. Del Giudice. *Bistable perception modeled as competing stochastic integrations at two levels. PLoS Comput Biol*, 5 (2009) e1000430.

[16] W. Weaver. *Science and complexity*. Am. Sci., 36 (1948) 536.

[17] K. Amunts, C. Ebell, J. Muller, M. Telefont, A. Knoll, and T. Lippert. *The Human Brain Project: Creating a European Research Infrastructure to Decode the Human Brain*. Neuron, 92 (2016) 574.

Paolo Del Giudice: è Primo Ricercatore dell'Istituto Superiore di Sanità a Roma e Professore incaricato di Reti Neurali presso il Dipartimento di Fisica dell'Università di Roma3. Si occupa dello studio teorico di modelli neuronali e della loro inferenza dai dati elettrofisiologici.

Maurizio Mattia: è Ricercatore dell'Istituto Superiore di Sanità a Roma e Professore incaricato di Neural Networks presso il Dipartimento di Fisica dell'Università di Roma 'Sapienza'. Si occupa dello studio teorico delle reti neuronali e dell'analisi dell'attività nervosa in collaborazione con laboratori sperimentali nell'ambito dello 'Human Brain Project' (finanziamento "EU H2020 Research and Innovation", grant n. 945539 - HBP SGA3).

L'elaborazione d'informazione nelle reti neurali

If a machine is expected to be infallible, it cannot also be intelligent.

A. Turing

Elena Agliari *Dipartimento di Matematica "Guido Castelnuovo" – Sapienza Università di Roma.*

Adriano Barra *Dipartimento di Matematica & Fisica "Ennio De Giorgi" – Università del Salento.*

L'intento di queste note è mostrare come due *modi operandi* tipici, rispettivamente, della Matematica (l'inferenza statistica) e della Fisica (la meccanica statistica) possano fungere da pilastri concettuali sui quali erigere una *teoria delle reti neurali*, a dire, un telaio logico-deduttivo nel quale rappresentare le reti di neuroni e dal quale evincere, come proprietà emergenti delle stesse, le capacità intellettive superiori di cui queste fruiscono. Le abbiamo volutamente chiamate emergenti, poiché, come mostreremo nella prima sezione, il singolo neurone può essere relegato ad un semplice interruttore, un sommatore a soglia rumoroso, lontano dal manifestare le suddette capacità che quindi scaturiscono dalla rete in quanto tale e non dai suoi singoli costituenti elementali. Nel resto delle scritto discuteremo un *paradigma teorico minimale*: divideremo il processo della *cognizione* in due momenti, quello dell'*apprendimento* ("learning") e quello dell'*impiego* di ciò che si è appreso ("retrieval") e mostreremo come l'inferenza statistica sia il linguaggio naturale per il primo momento mentre la meccanica statistica lo sia per il secondo (ed il confine tra le due alquanto sfumato). Per descrivere questi processi sfrutteremo due modelli paradigmatici per l'apprendimento automatico artificiale e per le reti neurali biologiche, ovvero, rispettivamente, la macchina di Boltzmann e la rete di Hopfield. Infine, nella chiusura dello scritto, mostreremo come, dal punto di vista astratto della processazione d'informazione, questi modelli siano due facce di un'unica medaglia, rendendo apprendimento ed impiego (d'informazione) un tutt'uno, i.e. il fenomeno cognitivo.

L'intersezione tra le Neuroscienze e l'Intelligenza Artificiale

Il modo in cui il cervello rappresenta ed elabora le informazioni sul mondo, l'emergenza della coscienza e la potenziale ricaduta applicativa delle macchine intelligenti sono oggigiorno tra i temi più caldi della Scienza. Recenti progressi nella comprensione del funzionamento del cervello (e.g., attraverso registrazioni multielettrodo per sondare l'attività cerebrale) e nel miglioramento delle prestazioni della sua controparte sintetica (e.g., attraverso nuove tecnologie come le GPU, la creazione di enormi database e lo sviluppo di algoritmi per l'elaborazione di questi *big data*) hanno suscitato profondo interesse e clamore. Tra gli scopi di questo articolo divulgativo è anche l'apprezzare come la modellistica matematica[1] – che è sempre stata soggiaciente tanto alle investigazioni nelle neuroscienze quanto in intelligenza artificiale (IA) – abbia eretto paradigmi validi in entrambe le declinazioni della processazione d'informazione e sia stata di fatto il loro *trait d'union* sin dalla genesi di queste discipline. In effetti, le neuroscienze e l'IA sono finalmente abbastanza mature da interagire ed autosostenersi reciprocamente, promuovendo ulteriormente il loro sviluppo: iniziative come lo Human Brain Project in Europa, il progetto BRAIN negli Stati Uniti, il programma Brain-MINDS in Giappone ed il China Brain Project hanno dimostrato l'entusiasmo e l'impegno dell'intera comunità scientifica su scala globale, tuttavia, l'interesse nel decifrare il *codice neurale* non è nato con quest'ultima ondata di entusiasmo, alimentata, di fatto, da progressi tecnologici più che concettuali. In effetti, come l'era moderna delle neuroscienze ha le sue radici nel metodo rivoluzionario di Golgi per colorare i neuriti e nelle indagini pioneristiche di Cajal perpetrate più di un secolo fa, alla stessa stregua l'IA pone le sue fondamenta nella macchina di Babbage (primordiale rotativa da calcolo derivata dal telaio tessile), in ultima istanza frutto della rivoluzione industriale inglese avvenuta oltre due secoli orsono.

In estrema sintesi, e restringendo il nostro discorso dalla metà del Novecento ai nostri giorni (dove si concentrano i principali risultati), una volta effettuati i primi esperimenti sulla comunicazione elettrica tra neuroni, immediatamente modelli matematici che mimassero l'emissione di impulsi elettrici analoghi a quelli sperimentalmente rivelati iniziarono a proliferare (dai neuroni di Hodgkin e Huxley, a quelli di Stein, a quelli di Nakubo, etc.) fino a culminare nel *perceptrone* di Rosenblatt del 1958 dove si offriva una prima trattazione logica dei neuroni astratti e delle loro capacità (suggerendone un impiego artificiale). Questa prima ondata di entusiasmo frenò bruscamente nel 1969 con l'uscita del libro *Perceptron* di Minsky e Papert, i quali mostrarono come questi modelli matematici di neuroni per la computazione spontanea fossero inadatti a risolvere perfino banali operazioni di logica elementare (e.g., tecnicamente si fermavano allo *XOR* perché erano intrinsecamente classificatori linari): questa doccia fredda, che fece sprofondare quest'intersezione tra le due Scienze in quello che è chiamato *the winter time* nella Comunità, fu in realtà una felice cornucopia poiché (tenendo a mente che nel mentre una meccanica statistica complessa per i modelli di campo medio era ormai a buon punto [1]) spostò l'attenzione dal singolo neurone alle reti di neuroni. Prendendo spunto dalla geniale idea di Hebb sull'apprendimento sinaptico, nel 1982 Hopfield [2] – ed indipendentemente Little e Amari – sviluppò un modello minimale di rete neurale che aveva capacità emergenti di gran lunga superiori a quelle che i singoli neuroni, per quanto *complessificati*, riuscissero ad avere. La rete di Hopfield è un grafo completamente connesso sui cui nodi vivono dei neuroni binari (on/off) ed i cui archi mimano le connesioni sinaptiche tra gli stessi e possono essere associati a pesi (efficacie sinaptiche) sia positivi che negativi: dal punto di vista della meccanica statistica questo sistema complesso è un vetro di spin (si veda, per una definizione di vetro di spin, la *lezione mancata*). Come mostreremo, questo sistema presenta comportamenti non banali che naturalmente elessero la meccanica statistica a disciplina cardine per lo studio teorico di questi modelli di reti neurali. In particolare,

[1] Nella *lezione mancata* di questo numero di Ithaca dedicato all'Intelligenza Artificiale approfondiamo alcune basi teoriche necessarie per una migliore comprensione della modellistica affrontata nel presente lavoro divulgativo (ed in molti altri di questo volume): in particolare, tutte le Hamiltoniane usate in questo articolo sono forme quadratiche.

tra il 1979 ed il 1980, Parisi [3] mise in luce nei vetri di spin proprietà ultrametriche che, da un punto di vista cognitivo, si è tentati associare alla categorizzazione genere-specie che spontaneamente tendiamo a fare. Qualche anno più tardi, nel 1985 Amit, Gutfreund e Sompolinsky [4] studiarono il modello di Hopfield sfruttando idee e tecniche meccanico-statstiche originariamente sviluppate per l'indagine dei vetri di spin ottenendo la prima trattazione sistemica interamente meccanico statistica di una rete neurale[2].
In queste note informali ripercorreremo in primis la via di Hopfield, ispirato dalle reti biologiche, per poi approdare a modelli di impiego nella controparte artificiale per infine mostrare come le due vie, la biologica e l'artificiale, siano coincidenti dal punto di vista astratto della processazione di informazione mediante generiche reti frustrate[3].

Dinamica di singolo neurone

In questa sezione prenderemo confidenza con gli attori principali del nostro cervello, i *neuroni* (i nodi della rete neurale) e le loro connessioni, dette *sinapsi* (gli archi della rete)[4]. Queste componenti evolvono su scale così distanti che, come sovente accade in Fisica (e.g., nell'approssimazione di Born-Oppeneimer in struttura della materia o in tutta la termodinamica adiabatica), possiamo trattare separatamente le relative dinamiche: quando ci preoccuperemo dei neuroni

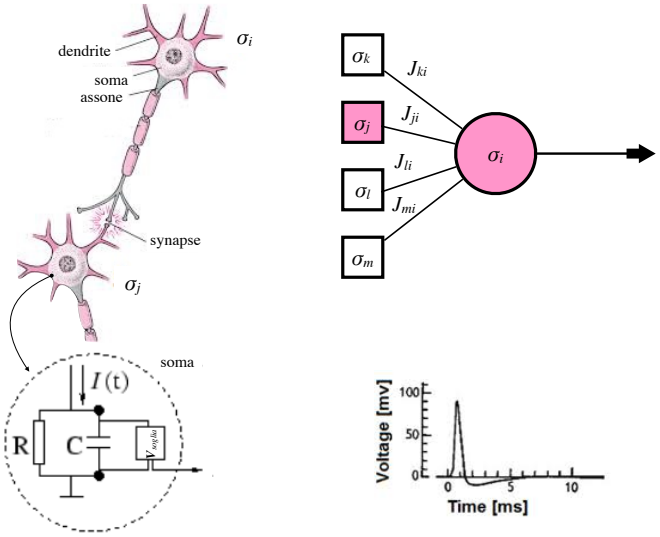


Figura 1: *Rappresentazione schematica di un neurone biologico (a sinistra) e di un neuron artificiale/logico (a destra). Nel primo sono mostrati alcuni dettagli biologici (assone, dendriti, soma ed un contatto sinaptico con un neurone afferente dal basso) ed il loro* equivalente circuitale *come integratore RC, mentre nel secondo si vede come questo sommi gli stimoli provenienti dai neuroni adiacenti, pesandoli mediante il filtro sinaptico J, ed eventualmente producendo un segnale a sua volta (in basso a destra mostriamo inoltre l'andamento temporale tipico di un singolo spike).*

considereremo le sinapsi *congelate* (come gli accoppiamenti nei vetri di spin), viceversa quando ci interesseremo alla dinamica sinaptica potremo tralasciare l'influenza di quella neurale (poiché lo stato dei neuroni potrá essere *mediato via*).
Un'altra similitudine con la Fisica, anticipata nella sezione precedente, risiede nel fatto che tanto i neuroni possono essere (dal punto di vista della processazione di informazione) fondamentalmente in due stati ("on", emettono un segnale elettrico ed "off", rimangono quiescienti) alla stregua degli spin di Ising, tanto le sinapsi possono essere sia eccitatorie (mimando gli accoppiamenti positivi) che inibitorie (mimando quelli negativi), in ultima istanza permettendoci così di sancire che, da una prospettiva meccanico statistica, la rete neurale è un vetro di spin.
A seguire presenteremo due modelli stilizzati di neurone, il primo (il modello integrate-and-fire di Stein) ambisce a fornire un supporto ai fisiologi alle prese con le reti neurali biologiche, mentre il secondo (il modello logico di McCulloch e Pitts) costituisce un tassello fondamentale

[2] L'impiego della meccanica statistica offriva un ulteriore vantaggio pratico in quanto non richiede una descrizione particolarmente dettagliata dei componenti del sistema oggetto di studio (e.g., non è necessaria una conoscenza minuziosa della posizione e della velocità di ciascuna particella per sviluppare un modello di gas per determinarne pressione e temperatura) ed effettivamente negli anni '80 le informazioni disponibili sulla struttura microscopica delle reti neurali biologiche erano ancora piuttosto limitate.

[3] Per la comprensione, cruciale, dell'aggettivo *frustrato* si veda di nuovo la *lezione mancata*.

[4] Più precisamente, un neurone è costituito da un soma, i.e., il corpo cellulare, da un "cavo di uscita dove eventualmente propagare il segnale elettrico" chiamato *assone* e da molti "cavi di entrata", che formano l'*albero dendritico* al quale neuroni afferenti mandano i loro stimoli: l'assone di un neurone afferente ed il dendrite del neurone ricevente sono connessi mediante le sinapsi: dal punto di vista della processazione d'informazione, i costituenti passivi (i.e., assoni e dendriti) non giocano un ruolo saliente e verranno perciò trascurati per semplicità.

nella controparte artificiale.

- Neurone (biologico) di Stein
  Nel neurone di Stein gli elementi cardine sono la differenza di potenziale della membrana esterna del neurone (a dire il voltaggio che si registra mettendo un elettrodo all'interno del soma ed uno all'esterno), la sua resistenza R, la sua capacità C, le correnti afferenti da neuroni esterni e la possibilità di generare a sua volta una corrente (un segnale elettrico impulsato che viene chiamato *spike*[5]), si veda la figura 1. Il tutto si amalgama nell'equazione differenziale per l'evoluzione del potenziale di membrana V_i per il neurone i-esimo che si legge

$$\frac{dV_i}{dt} = -\frac{V_i}{\tau} + \sum_{j\neq i}^{N} J_{ij} \sum_{k}^{T} \delta(t - t_{kj} - d_{ij}), \quad (1)$$

  dove $\tau := RC$ funge da costante di tempo di questo "circuito integratore", mentre la corrente afferente al neurone in esame è costituita dalla somma (lineare) dei contributi provenienti dai vari neuroni ad esso connesso, ognuno pesato mediante l'efficacia sinaptica J_{ij} e ricevuto stocasticamente a tempi diversi t_{kj} e ritardati dalla propagazione stessa mediante i d_{ij}. Questa dinamica neurale è dissipativa, in virtù del termine $-V_i$, ma continuamente rinvigorita da stimoli esterni: poiché le sinapsi possono sia aumentare la differenza di potenziale del neurone afferente (sinapsi eccitatorie) sia diminuirla (sinapsi inibitoria), l'evoluzione del potenziale sinaptico compie un moto erratico e se questo raggiunge una soglia critica per la stabilità della membrana, semplificando oltremodo, questa "si rompe temporaneamente", dando luogo ad un impulso elettrico, i.e. lo spike, rivolto ai neuroni riceventi (ognuno dei quali lo peserà, in concerto con altri afferenti da neuroni terzi, positivamente o negativamente in ragione della sinapsi che congiunge l'assone di output con i dendriti di input)[6]. Un esempio della dinamica neurale fino all'emissione di uno spike è mostrato in figura 2.

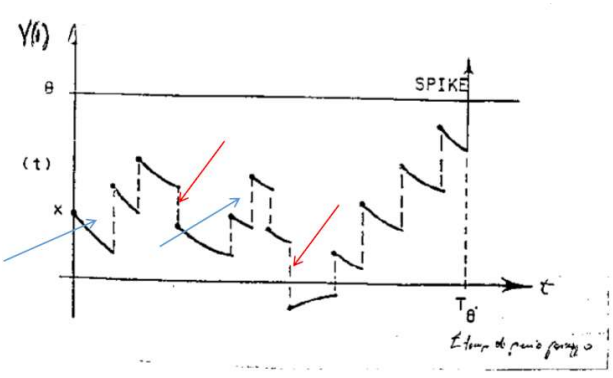


Figura 2: *Esempio dell'evoluzione temporale erratica del potenziale di membrana di un neurone fino alla sua generazione dello spike. I salti discontinui sono dovuti a spikes provenienti da neuroni afferenti (in blu filtrati da sinapsi eccitatorie ed in rosso da sinapsi inibitorie).*
Nota: in basso si legge in corsivo "tempo di primo passaggio": è la scrittura di Daniel Amit, pioniere delle reti neurali (l'immagine e' presa dal suo corso di Reti Neurali, tenuto in Sapienza dagli anni novanta fino al 2005).

- Neurone (artificiale) di McCulloch&Pitts
  In questo neurone logico, si veda la figura 1, si trascurano i dettagli fisici (quali le dissipazioni ed i ritardi di propagazione) e si mette il fuoco solamente sulle capacità di calcolo dello stesso. Usando gli stessi simboli del neurone di Stein possiamo scrivere

$$V_i(t + \Delta t) = \Theta\left(\sum_{j\neq i}^{N} J_{ij} I_j - V_{\text{soglia}}\right), \quad (2)$$

  dove Θ è la funzione di Heaviside: lo stato di uscita del neurone (i.e., il potenziale della sua membrana) è fisso a zero a meno che la somma dei contributi elettrici afferenti non superi una soglia V_{soglia}, in qual caso il neurone emette il *potenziale d'azione*, cosa che si avverte notando che lo stato di uscita del neurone diventa uno. Nel neurone artifi-

[5] La genesi dello spike è dovuta ad un brusco crollo della stabilità della membrana cellulare, la quale, se destabilizzata dai continui spikes a sua volta ricevuti, si apre per dar luogo a sua volta al prosieguo della comunicazione nervosa lungo l'assone della cellula in questione, alla volta degli alberi dendritici di altri neuroni con cui questo è connesso.

[6] Per un approfondimento sulle reti neurali biologiche si veda il contributo di Paolo Del Giudice & Maurizio Mattia in questo volume.

ciale si possono quindi rappresentare i due stati logici di Boole[7].

Immaginando di scrivere ora l'evoluzione del potenziale di Stein, l'eq. 1, o di quello di McCulloch&Pitts, l'eq. 2, non più per il singolo neurone i-simo, ma per tutti gli N neuroni che compongono la rete (risultando quindi in un sistema di N equazioni differenziali accoppiate), la presenza sottostante di una rete di neuroni interconnessi appare nitida, alla pari delle piuttosto limitate capacità di processare l'informazione da parte del singolo neurone: l'atto di cognizione deve emergere come un fenomeno collettivo della rete (motivo per cui, a seguire, chiamiamo in causa la Meccanica Statistica come modus operandi per investigarlo).

La cognizione nelle reti neurali

Come abbiamo visto nella precedente sezione, ogni singolo neurone può vivere (in prima approssimazione) in due stati: quiesciente o emettitore di segnale. Prendiamo a prestito dalla Fisica lo spin di Ising $\sigma_i \in \{-1, +1\}$ per caratterizzarlo in maniera tale che, finché V_i è minore della soglia, si ha $\sigma_i = -1$, mentre quando V_i raggiunge la soglia per l'emissione dello spike $\sigma_i \to +1$. Consideriamo ora una rete costituita da N neuroni $\{\sigma_i\}_{i=1,...,N}$ e scriviamo la legge evolutiva per il generico neurone i-esimo come

$$\sigma_i(t+1) = \text{sign}[\tanh(\beta h_i(t)) + \eta_i(t)], \quad (3)$$

$$h_i(t) = \sum_{j \neq i}^{N} J_{ij}\sigma_j(t) + h_i^{ext}, \quad (4)$$

dove h_i è il campo afferente sul neurone σ_i in esame (ed è costituito dalla somma lineare di tutti gli stimoli prodotti dai neuroni afferenti $\{\sigma_j\}_{j\neq i}$, pesati con le rispettive efficacie sinaptiche J_{ij} e da un eventuale stimolo esterno h^{ext}) mentre la tangente iperbolica rilassa l'assunto di assenza di rumore tacito nella funzione a gradino di McCullch&Pitts: η_i è una variabile casuale uniformemente distribuita in $[-1, +1]$ e $\beta \in \mathbb{R}^+$ è un parametro che modula la stocasticità del processo in maniera tale che quando $\beta \to \infty$ la dinamica è deterministica e lo spin si allinea alla direzione del campo h_i (e si riottiene il comportamento logico di McCulloch&Pitts); quando $\beta \to 0$ i campi diventano impercettibili e la serie temporale degli stati neurali diventa una sequenza di Bernoulli di questionabile interesse per gli scopi di questo scritto. In un contesto meccanico-statistico classico β gioca il ruolo dell'inverso della temperatura (in unità opportune)[8].
Questa dinamica si può scrivere, per l'intera rete, in termini probabilistici, introducendo la probabilita' $P_t(\sigma)$ di trovare, al passo di aggiornamento t, la rete in un generico stato $\sigma := \{\sigma_1, ..., \sigma_N\}$, tra i 2^N possibili, come

$$P_{t+1}(\sigma) = \prod_{i=1}^{N} \frac{e^{\beta\sigma_i h_i(\sigma(t))}}{2\cosh[\beta\sigma_i h_i(\sigma(t))]},$$

alla volta di un processo di Markov

$$P_{t+1}(\sigma) = \sum_{\sigma'} W[\sigma;\sigma']P_t(\sigma'),$$

con W opportuna matrice di transizione. È possibile dimostrare che questo processo è ergodico[9] e, per $t \to \infty$, $P_t(\sigma)$ converge ad un'unica distribuzione stazionaria. Inoltre, se ci restringiamo a considerare efficacie sinaptiche simmetriche (i.e., $J_{ij} = J_{ji}$), la dinamica soddisfa il *bilancio dettagliato*, il quale garantisce che lo stato stazionario a cui il sistema rilassa sia uno stato di equilibrio e la relativa distribuzione abbia la forma funzionale della distribuzione di Gibbs

$$\lim_{t\to\infty} P_t(\sigma) =: P(\sigma) \propto \exp[-\beta H(\sigma|J)] \quad (5)$$

per qualche opportuna funzione costo $H(\sigma|J)$ (o Hamiltoniana se si vuole preservare il gergo fisico). Questa informazione, come vedremo nelle due prossime sezioni, è cruciale tanto per l'apprendimento quanto per l'impiego di ciò che si è appreso. Nel seguito, per chiarezza espositiva, tratteremo prima il richiamo alla memoria di informazione precedentemente appresa e dopo ci concentreremo sul processo di apprendimento[10].

[7] Per un approfondimento sulle reti neurali artificiali si veda il contributo di Giorgio Buttazzo in questo volume.

[8] Si veda a questo proposito il riquadro "La temperatura ubriaca" nella *lezione mancata*.

[9] L'ergodicità vale quasi ovunque in β, ovvero ad eccezione del limite $\beta \to \infty$; per maggiori dettagli rimandiamo a [5, 6].

[10] Si veda anche il contributo di Daniele Tantari in questo volume per un simile approccio.

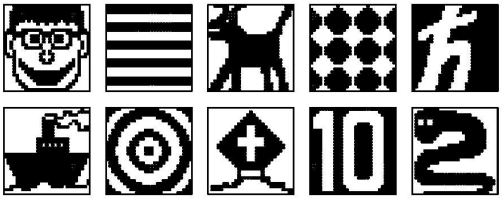


Figura 3: *Esempi di dieci immagini in bianco e nero formate da 30×30 pixel, incamerate da una rete di Hopfield di 900 neuroni binari.*

Il paradigma minimale del retrieval: meccanica statistica

Alla volta di un'Hamiltoniana efficace da inserire nell'esponenziale di Gibbs nella distribuzione(5), introdotti gli N neuroni di Ising $\{\sigma_i\}_{i=1,...,N}$, dobbiamo specificare meglio la matrice sinaptica J_{ij}. Per fare questo introduciamo il generico concetto di *pattern*, ξ che non è altro che un'informazione codificata in un linguaggio binario: un pattern può essere un concetto, una parola, un'immagine, etc.. Qui assumeremo che un pattern rappresenti un'immagine in bianco e nero, codificata attraverso una stringa di lunghezza fissa, costituita da N bit (si veda figura 3).
Per semplicità (ma non solo[11]) lavoreremo solo con patterns random: un pattern $\xi = (\xi_1, \xi_2, ..., \xi_N) \in \{-1, +1, \}^N$ si genera estraendo l'elemento ξ_i secondo la probabilità $P(\xi_i = +1) = P(\xi_i = -1) = 1/2$, per ogni $i \in (1, ..., N)$. Inoltre, poiché non vogliamo usare un'intera rete neurale (cioè N neuroni) per gestire un unico pattern ξ, ma per gestirne $P \sim O(N)$, distingueremo i vari pattern attraverso un'etichetta: $\xi \to \xi^\mu$, $\mu \in (1, ..., P)$. Siccome sappiamo che una qualunque (ragionevole) dinamica neurale stocastica converge necessariamente verso le configurazioni corrispondenti ai minimi della funzione costo $H(\sigma|J)$, il punto fondamentale ora è *incastonare* questi pattern nei minimi della funzione costo in maniera tale che la termalizzazione del sistema (garantita dal bilancio dettagliato) faccia evolvere lo stato neuronale da una configurazione iniziale $\sigma(0)$ ad una configurazione $\sigma(t) = \xi^\mu$ stabile per tutti i tempi successivi (almeno entro un certo grado di errore); questo fenomeno viene interpretato come la ricostruzione del pattern μ-esimo[12]. Il particolare pattern ξ^μ a cui il sistema spontaneamente rilassa dipenderà dallo stato iniziale $\sigma(0)$, interpretato come input della rete. Detto in altre parole, vogliamo definire la matrice delle interazioni J in modo che le configurazioni neurali corrispondenti a ciascuno dei P pattern costituiscano degli attrattori, si vedano le figure 4 e 5. A questo scopo, la scelta più intuitiva è $H(\sigma|J) \sim -N^{-1} \sum_{\mu=1}^{P} (\xi^\mu \cdot \sigma)^2$, infatti, se la rete

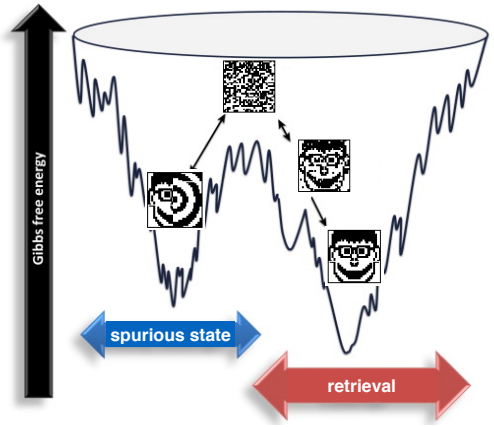


Figura 4: *Rappresentazione grafica dell'energia (libera) del modello di Hopfield. Nel regime in cui la rete lavora al meglio, gli stati corrispondenti ai pattern sono minimi globali e, poiché le dinamiche stocastiche a cui ogni neurone obbedisce devono convergere ad essi, la termalizzazione nella meccanica statistica coincide con il riconoscimento dei pattern mediante memoria associativa nella teoria delle reti neurali. Il profilo energetico presenta anche minimi locali, ovvero stati mestastabili, che tipicamente corrispondono a combinazioni degli stessi pattern memorizzati, anche detti stati spuri; il rilassamento ad uno di questi stati viene considerato come un errore da parte della memoria.*

[11] Si potrebbe obiettare – a ragione [7] – che una teoria random non abbia granché senso come *oscillatore armonico* di una teoria per le reti neurali. Osserviamo però che per il teorema di compressione di Shannon se la rete in esame è in grado di gestire P patterns casuali sarà certamente in grado di gestirne almeno lo stesso numero se correlati o con struttura al loro interno: la teoria random può fornire utili limiti ed offre un classico quadro di riferimento (dove tutto fattorizza asintoticamente) [8]. Di contro è parimenti d'obbligo, convenire che molti dei problemi interessanti in IA sono proprio legati alla presenza di struttura nei dataset (si veda a tal proposito il contributo di Matteo Marsili in questo volume e si approcciano coerentemente con telai inferenziali *profondi*, i.e a molti strati [9], per i quali si vedano i contributi di Guido Sanguinetti e Carlo Lucibello nel presente volume.)

[12] Si pensi per esempio come immagine ad un volto a noi ben noto: se ci viene presentato solo un dettaglio, per esempio gli occhi, subito noi sappiamo riconoscere, ovvero riportare alla memoria (fare *pattern recognition* mediante memoria associativa) l'intero volto.

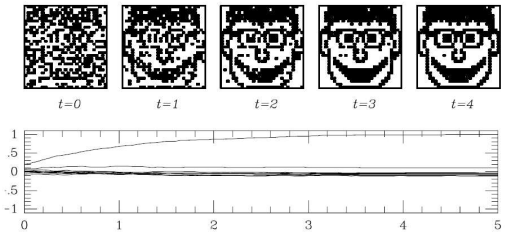


Figura 5: *Richiamo di un pattern. Al crescere del suo parametro d'ordine da zero verso uno (cioè al termalizzare della rete), si vede il ricostruirsi della* faccia *che via via appare alla memoria. Parimenti sotto viene mostrata l'evoluzione delle* 10 *magnetizzazioni di Mattis (una per pattern incamerato) e si osserva come una -quella relativa al pattern richiamato- converga monotonamente al suo massimo mentre le altre nove oscillino (con intensità $\sim N^{-1/2}$) intorno allo zero.*

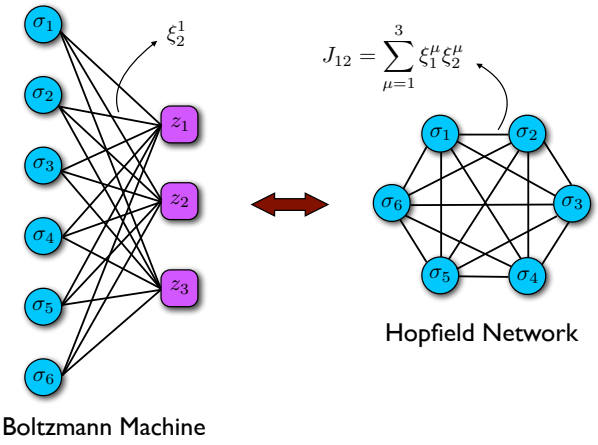


Figura 6: *Rappresenzazioni schematiche di una macchina di Boltzmann (sinistra), archetipo della macchina che apprende in machine learning, e di una rete di Hopfield (destra), oscillatore armonico delle reti biologiche, i.e. memorie associative che eseguono pattern recognition.*

non riconosce alcun pattern, l'energia è circa nulla, mentre se riconosce un certo pattern la rete acquista un'energia $-\mathcal{O}(N)$, di gran lunga più conveniente[13]. Il peso sinaptico tra il neurone i ed il neurone j risulta pertanto $J_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$: si implementa in maniera naturale l'idea di apprendimento di Donald Hebb (erede di Ivan Pavlov), per cui *neuroni che emettono congiuntamente e reciprocamente segnali elettrici rinforzano i relativi canali di comunicazione (qui schematizzati nella matrice sinaptica)*[14] [10]. Possiamo quindi scrivere l'Hamiltoniana di Hopfield come

$$H_H(\sigma|J) = -\frac{1}{2N}\sum_{i,j}^{N,N}\sum_{\mu=1}^{P}(\xi_i^\mu \xi_j^\mu)\sigma_i\sigma_j \quad (6)$$

$$= -\frac{N}{2}\sum_{\mu=1}^{P} m_\mu^2, \quad (7)$$

dove il pedice H nell'Hamiltoniana sta per Hopfield e nella seconda riga abbiamo introdotto i P parametri d'ordine *magnetizzazioni di Mattis*, definiti come $m_\mu := N^{-1}\sum_{i=1}^{N}\xi_i^\mu\sigma_i$ (per avere una rappresentazione grafica della rete si veda Figura 6, grafo di destra). A questo punto è elementare notare che, al fine di minimizzare l'energia H_H alla rete convenga avere una[15] magnetizzazione di Mattis pari ad uno: ai neuroni della rete, per vivere comodi, conviene organizzarsi e questa loro organizzazione spontanea produce proprio il richiamo alla memoria di patterns precedentemente appresi.
Lo schema di apprendimento Hebbiano genera nel profilo energetico non solo i P minimi globali corrispondenti ai pattern memorizzati, ma anche un grandissimo numero (esponenzialmente crescente in N) di minimi locali tipicamente corrispondenti a "misture" di pattern, anche detti stati spuri (e.g., lo stato $\sigma_i = \mathrm{sgn}(\xi_i^1+\xi_i^2+\xi_i^3)$, per $i = 1, ..., N$). Quando il rapporto tra il numero di pattern ed il numero di neuroni è troppo alto (i.e., $\alpha = P/N > \alpha_c \approx 0.14$), questi minimi locali dominano il panoramo energetico ed il sistema non è più in grado di richiamare correttamente. Da queste considerazioni emerge anche che, a nostro avviso, questo modello debba rientrare nei cosiddetti *sistemi complessi*: questo non stupisce poichè,

[13] Osserviamo che, per costruzione, i pattern $\{\xi^\mu\}_{\mu=1,...,P}$ sono tra loro ortogonali (almeno nel limite di N grande), ovvero $\xi^\mu \cdot \xi^\nu = 0$, per ogni $\mu \neq \nu$. Di conseguenza, la ricostruzione di un pattern, diciamo ξ^ν, comporta che lo stato neuronale sia (circa) ortogonale a tutti gli altri pattern e di conseguenza l'unico contributo non nullo all'energia $H(\sigma|J)$ proviene dal termine ν-esimo della sommatoria.

[14] Questa idea è molto semplice ed è di fatto *economia idraulica*: se il neurone i sta emettendo incessantemente segnali al neurone j e viceversa, mentre non ne sta mandando al neurone k, conviene ampliare il canale di comunicazione tra i e j e diminuire quello tra i e k per preservare l'omeostasi della rete e minimizzare la congestione di segnali, proprio come si cerca di minimizzare le impedenze in una rete idraulica.

[15] Come ricordato prima, poiché i pattern sono ortogonali, la rete può richiamare correttamente solo una memoria per volta.

essendo le sinapsi (cioè gli accoppiamenti tra i neuroni) tanto eccitatorie (i.e. positive) quanto inibitorie (i.e. negative), il modello di Hopfield è, dal punto di vista della Meccanica Statistica, un particolare esempio (molto interessante) di spin glass [16]. Dal punto di vista statistico questo modello cattura e riproduce fedelmente funzioni di correlazione a due punti, che forniscono una sufficiente caratterizzazione per pattern randomici appartenenti allo scenario classico descritto dai Teoremi del Limite Centrale. Se volessimo cimentarci in problemi molto più complessi (e.g., la comprensione del linguaggio naturale, la comprensione dei dati generati al CERN per lo studio della fisica delle alte energie o il tracciamento automatico per il contenimento di una pandemia[17]) probabilmente dovremmo optare per funzioni costo in grado di inferire bene funzioni di correlazione a molti punti: questo risulterebbe in un telaio inferenziale lontano da una distribuzione di Gibbs di un'Hamiltoniana quadratica: molti dei progressi di cui l'IA si fa artefice in questi anni avvengono mediante architetture *deep* [9] (che non affrontiamo in questo articolo).

Il paradigma minimale del learning: inferenza statistica

In questa sezione dedicata all'apprendimento inizieremo descrivendo un classico modello di rete neurale artificiale, per poi mostrare come, in ultima istanza, questa via non sia altro che una rilettura del paradigma Hebbiano, opportunamente declinato per le macchine piuttosto che per la materia organica.
Sempre restringendoci a reti per le quali valga il bilancio dettagliato[18], uno dei mattoni fondamentali sui quali si erigono oggi architetture da calcolo per il moderno *deep learning* è certamente la Macchina di Boltzmann (*Boltzmann machine* [14, 15]): questa è una rete bipartita, con uno strato di neuroni visibile, a cui viene fornito input dall'esterno, ed uno strato di neuroni nascosto, atto a scovarne le correlazioni (si veda figura 6, grafo di sinistra).
Nello specifico assumiamo che lo strato di neuroni visibile sia composto da N spin di Ising $\sigma_i = \pm 1, i \in (1, ..., N)$, mentre lo strato nascosto sia composto da P spin di Ising[19] $\tau_\mu = \pm 1, \mu \in (1, ..., P)$, assumiamo inoltre che esistano delle connessioni sinaptiche unicamente tra neuroni di strati diversi, quindi la matrice sinaptica sia una matrice $N \times P$ il cui ingresso generico chiamiamo $\xi_i^\mu \in \mathbb{R}$ con $i \in (1, ..., N)$ e $\mu \in (1, ..., P)$. Dal punto di vista della meccanica statistica, questa macchina null'altro è che uno spin glass bipartito, per cui la funzione costo (o Hamiltoniana) H_B, dove il pedice ricorda il nome Boltzmann, e relativa misura di Gibbs, si scrivono rispettivamente

$$H_B(\sigma, \tau|\xi) = \frac{-1}{\sqrt{N}} \sum_{i,\mu}^{N,P} \xi_i^\mu \sigma_i \tau_\mu - \sum_i^N h_i^{ext} \sigma_i \quad (8)$$

$$P_\xi(\sigma, \tau) = \frac{e^{-\beta H_B(\sigma,\tau|\xi)}}{Z_B}. \quad (9)$$

dove $\mathbf{h}^{ext}$ è un campo esterno al fine di far interagire la rete con il mondo esterno e $Z_B := \sum_\sigma^{2^N} \sum_\tau^{2^P} e^{-\beta H_B(\sigma,\tau|\xi)}$ è un fattore di normalizzazione, chiamato *funzione di partizione* in Meccanica Statistica[20].
Mentre il processo di richiamo coinvolge la dinamica neurale e fonda le sue radici teoriche nella Meccanica Statistica (in ultima istanza nell'estremizzazione dell'entropia vincolata da opportune funzioni costo, per dirla à la Jaynes), il processo di apprendimento coinvolge la dinamica sinaptica e fonda le sue radici teoriche nell'Inferenza

[16] In relazione alla sezione sul *Riduzionismo Statistico* nella Lezione Mancata, la rete di Hopfield ha una funzione costo quadratica (inevitabilmente, poichè abbiamo richiesto un minimo parabolico nel caso di correlazione tra lo stato della rete ed il pattern, i.e., $H_H \sim -(\sigma \cdot \xi)^2$) ed abbiamo visto come questo risulti in sostanziale accordo con le osservazioni empiriche di Pavlov ed Hebb sulle caratteristiche che la matrice sinaptica debba presenzare (la regola d'apprendimento di Hebb).

[17] Questi tre temi sono trattati, rispettivamente, nei contributi di Valerio Basile, Konstantinos Bachas & Alfredo Braunstein, Luca Dall'Asta ed Alessandro Ingrosso nel presente volume.

[18] La richiesta della simmetria degli accoppiamenti ci permette di parlare di una fetta, comunque cospicua, di modelli ma esclude importanti macchine e relativi algoritmi di apprendimento, probabilmente prime tra tutte la *back-propagation* nelle reti *feed-forward*, il *reinforcement learning* e le *convolutional networks*.

[19] In questo articolo divulgativo non entreremo in tecnicismi, ma la scelta del tipo di neurone da usare influenza in maniera cruciale il funzionamento della macchina [12], si vedano a tal proposito i contributi di Aurèlienne Decelle e Carlo Lucibello nel presente volume.

[20] Ed è importante notare che nel tentare di calcolarla con forza bruta, tipicamente fallendo, ci si imbatte in un conto NP.

Statistica. In questo parallelo, facciamo una prima distinzione sostituendo all'entropia di Shannon della Meccanica Statistica, la mutua entropia $D(P|Q)$ di Kullback-Leibler che, introdotte due distribuzioni $P(\sigma)$ e $Q(\sigma)$ definite sullo stesso spazio di probabilità, ne misura la similarità e si scrive

$$D(P|Q) = \sum_{\sigma} P(\sigma) \ln \left(\frac{P(\sigma)}{Q(\sigma)} \right),$$

in maniera tale che quest'osservabile sia sempre non negativa ed uguale a zero se e soltanto se $P = Q$ quasi ovunque. Vogliamo usare questo strumento matematico, per esempio estremizzandolo opportunamente, per fare machine learning: la macchina di Boltzmann, assunto che abbia matrici sinaptiche simmetriche, ammette una rappresentazione probabilistica $P(\sigma, \tau)$ in termini di una distribuzione di Gibbs di un'opportuna funzione costo $H(\sigma, \tau|\xi)$. Sottilineiamo nuovamente che, in questo contesto, sono le ξ -le connessioni sinaptiche- i parametri liberi da poter variare a nostro vantaggio per far apprendere la rete (ovvero per l'estremizzazione della mutua entropia).
Sempre con fare pratico, per prendere confidenza, immaginiamo di avere M realizzazioni eventualmente rumorose (i.e., un campione casuale semplice) di uno dei previi pattern (cioè M immagini dello stesso soggetto), tutte di dimensione N (formate cioè da N pixel binari) che possiamo indicare come $\{\tilde{\sigma}_i^{(\alpha)}\}_{i=1,\dots,N}^{\alpha=1,\dots,M}$. Assumiamo che le M realizzazioni siano state generate identicamente ed indipendentemente dalla stessa distribuzione $Q(\tilde{\sigma})$. Ricordiamo che $Q(\tilde{\sigma})$ è incognita, ma possiamo stimarne i momenti a partire dagli esempi a disposizione, mentre $P_\xi(\sigma)$ ha solo la forma funzionale fissata (ed è una comoda famiglia di esponenziali peraltro). Scriviamo come punto di partenza la distribuzione congiunta $R(\tilde{\sigma}|\xi)$ dei dati e di $P_\xi(\sigma, \tau)$

$$\begin{aligned} R(\tilde{\sigma}|\xi) &= \prod_{\alpha=1}^{M} P_\xi(\tilde{\sigma}^{(\alpha)}, \tau) = e^{\sum_{\alpha=1}^{M} \ln P_\xi(\tilde{\sigma}^{(\alpha)}, \tau)} \\ &=: e^{\mathcal{L}(\xi|\tilde{\sigma}, \tau)}, \end{aligned} \tag{10}$$

quindi, invece di estremizzare $R(\tilde{\sigma}, \tau|\xi)$, possiamo usare $\mathcal{L}(\xi|\tilde{\sigma}, \tau)$, e parimenti non cambia nulla se alla stessa aggiungiamo e sottraiamo l'entropia empirica $S_M(Q) = -M^{-1} \sum_{\alpha=1}^{M} \ln Q(\tilde{\sigma}^{(\alpha)})$ ottenendo

$$\mathcal{L}(\xi|\tilde{\sigma}, \tau) = -\frac{1}{M} \sum_{\alpha=1}^{M} \ln \left(\frac{Q(\tilde{\sigma}^{(\alpha)})}{P_\xi(\sigma, \tau)} \right) - S_M(Q),$$

cioe' $\mathcal{L}(\xi|\tilde{\sigma}) = D_M(P|Q) - S(Q)$, dove la mutua entropia empirica si legge $D_M(P|Q) = \frac{1}{M} \sum_{\alpha=1}^{M} \ln \left(\frac{Q(\tilde{\sigma}^{(\alpha)})}{P_\xi(\sigma)} \right)$. La minimizzazione di $D_M(P|Q)$[21] ci porta a regole di apprendimento di sicura convergenza (nonostante nulla ci dica sui tempi per raggiungerla [13]).
Una volta applicata alla macchina di Boltzmann, nello scenario di apprendimento supervisionato (che ora introduciamo mediante il concetto di *media clamped*), questa procedura inferenziale risulta in una "ricetta" per l'apprendimento particolarmente intuitiva: introduciamo delle medie *clamped*, cioè dove lo strato visibile σ è costretto di volta in volta a *vedere* le M immagini, cioè tale che $P(\sigma) = \sum_{\tilde{\sigma}} P(\tilde{\sigma}) \delta(\sigma - \tilde{\sigma})$ (per esempio supplendo alla rete le varie immagini mediante il campo $\mathbf{h}$) e chiamiamo la media di Boltzmann di una generica osservabile O $\langle O \rangle =: \sum_{\sigma,\tau} O(\sigma, \tau) P_\xi(\sigma, \tau) / \sum_{\sigma,\tau} P_\xi(\sigma, \tau)$ mentre chiamiamo *clamped* la media con lo strato visibile forzato sull'immagine $\langle . \rangle_{clamped}$. La regola di apprendimento che otteniamo estremizzando l'entropia di Kullback-Leibler in questa maniera si legge

$$\Delta \xi_i^\mu = \epsilon \beta \left(\langle \sigma_i \tau_\mu \rangle - \langle \sigma_i \tau_\mu \rangle_{clamped} \right). \tag{11}$$

Quello che operativamente cerca di fare la macchina è quindi di riprodurre, operando in modalità autonoma, le correlazioni statistiche che la macchina vede, forzata a guardare il dataset.

Questo schema, sommariamente descritto, è chiamato *contrastive divergence* [14] ed è alla base di numerosi algoritmi di *machine learning*: a prima vista, poco sembra avere a che fare con le reti neurali di Hopfield ed il loro apprendimento Hebbiano di ispirazione biologica, ma non è così. Se infatti calcoliamo esplicitamente la funzione di partizione della macchina di Boltzmann

[21] e.g., mediante la regola del gradiente, cioè, chiamato ϵ un -piccolo- parametro di learning, $\xi(t + \epsilon) = \xi(t) - \epsilon \nabla_\xi D_M(P|Q)$.

marginalizzando sui neuroni τ scriviamo

$$
\begin{aligned}
Z_B &= \sum_{\sigma}^{2^N} \sum_{\tau}^{2^P} e^{\frac{\beta}{\sqrt{N}} \sum_{i,\mu}^{N,P} \xi_i^\mu \sigma_i \tau_\mu} \\
&= \sum_{\sigma}^{2^N} \prod_{\mu=1}^{P} 2\, e^{\ln\cosh(\frac{\beta}{\sqrt{N}} \sum_i^N \xi_i^\mu \sigma_i)} \\
&\sim \sum_{\sigma}^{2^N} e^{\frac{\beta}{2N} \sum_{i,j}^{N,N} \sum_\mu^P \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j} = Z_H . \quad (12)
\end{aligned}
$$

Come le due funzioni di partizione, quella della macchina di Boltzmann e quella della rete di Hopfield coincidono[22], così, con un semplice argomento Bayesiano si può infatti mostrare che per ogni dataset generato da un pattern, la macchina di Boltzmann fa evolvere i suoi pesi in maniera tale che, una volta espressa in termini duali di rete Hebbiana, la matrice sinaptica inferita abbia come ingresso generico J_{ij} proprio $J_{ij} = \hat{\xi}_i^\mu \hat{\xi}_j^\mu$, dove i cappucci rappresentano le medie campionarie (stimatori eccellenti) dei pixel dell'immagine contenuta nel dataset relativo al pattern μ: almeno all'interno di uno scenario elementare, alle prese con datasets e patterns randomici, c'è completa armonia nell'investigazione teorica tra reti neurali ispirate dalla biologia e reti neurali artificiali di impiego nel machine learning [16, 17].

Alcune considerazioni conclusive

Prima del cosiddetto *winter time*, venivano prodotti modelli matematici per i neuroni sia d'ausilio ai neurofisiologi per la comprensione dell'intelligenza biologica, sia come processori sintetici di informazione, alla volta dell'IA. Il modello di Hopfield per il retrieval (e la duale macchina di Boltzmann per il learning) discusso in questo articolo è nato con la fine del *winter time*: le critiche al perceptrone di Rosenblatt erano in realtà critiche all'analisi di singolo neurone ed hanno saggiamente spostato l'attenzione degli Scienziati dai neuroni alle reti di neuroni.

[22] E quindi anche la macchina di Boltzmann gode del diagramma di fase di Hopfield (si veda la figura 8), dove in questo caso $\alpha = P/N$ rappresenta il rapporto tra la grandezza dello strato nascosto e quella dello strato libero).

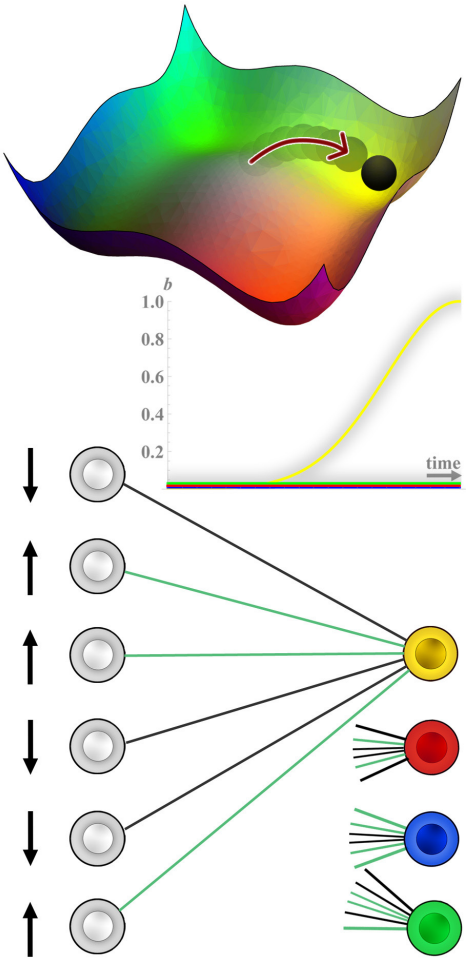


Figura 7: *Rappresentazione schematica di una macchina per il riconoscimento del colore: questa macchina di Boltzmann giocattolo è stata addestrata con la* contrastive divergence *a discriminare tra quattro colori. Di conseguenza, il panorama di energia (libera) della sua corrispettiva rete duale di Hopfield presenta quattro minimi (uno per ogni colore imparato). Se un colore viene in seguito presentato allo strato visibile della macchina (per esempio il* giallo*, rappresentato in un alfabeto binario fatto di spin sulla sinistra), i pesi che connettono questo strato a quello nascosto (se la macchina è stata addestrata con successo) sono tali da supplire al neurone nascosto la cui attivazione è responsabile dell'identificazione del giallo il campo massimo che lo strato visibile può produrre, lasciando i campi volti agli altri tre neuroni nascosti a valori minuscoli rispetto al primo. Intuitivamente si capisce anche perché ci sia un'α massimo in queste reti associative (si vedano le linee critiche nel diagramma di fase di Figura 8): chiaramente, tenendo fissato N, all'aumentare di P il numero di minimi in questo paesaggio deve continuare ad aumentare, ma in ultima istanza questi inizieranno a mischiarsi e fondersi uno nell'altro, rendendo il riconoscimento vacuo.*

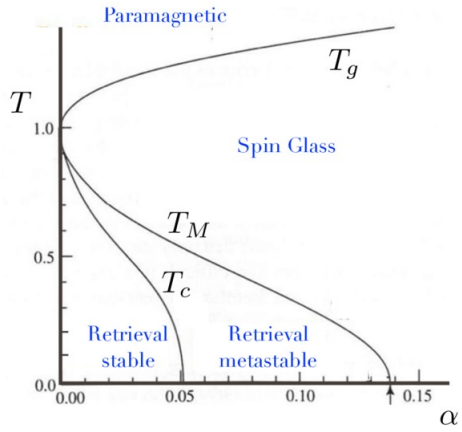


Figura 8: *Diagramma di fase della rete di Hopfield. Nel piano T, α – dove T rappresenta il rumore nella rete mentre $\alpha = P/N$ il suo carico (cioè il rapporto tra il numero di patterns che la rete deve gestire ed il numero di neuroni di cui essa e' composta) – troviamo diverse regioni, cioè diversi comportamenti della rete. Ad alte temperature, dove i neuroni non possono percepirsi reciprocamente, il sistema si comporta come un gas di spin, cioè un paramagnete ergodico. Abbassando la temperatura, partendo da carico nullo, la rete funziona bene come memoria associativa ed è in grado di riconoscere e distinguere tutti i P patterns fino alla prima linea critica, che sancisce il confine dello* stable retrieval*, dove si ha una transizione di fase e la rete entra in un nuovo regime dove i patterns sono ancora richiamabili, ma minimi spuri iniziano a dominare il paesaggio. Questo avviene in tutta la regione* retrieval metastable*, oltre la quale, per α ancora maggiori, la rete perde le sue capacità di pattern recognition e rimane uno spin glass senza proprietà di memoria.*

In questa era *post winter* nella quale viviamo circondati sempre più dall'IA, anche grazie alla dualità tra reti di Hopfield (archetipi dell'apprendimento biologico) e macchine di Boltzmann (oscillatori armonici dell'apprendimento artificiale) discussa in questo scritto, queste due branche della Scienza, che oggi si potrebbero chiamare Neurobiologia ed IA, si reincontrano (lavorando in connubio proprio come avvenne il secolo scorso alle prese con la dinamica di singolo neurone), ma a livello piu alto: questa volta sono le reti di neuroni ad interessare, le interazioni sopra i soggetti. Infatti, uno degli aspetti più significativi del modello di Hopfield è stato quello di spostare il focus dal singolo neurone alla rete [18]: la silente rivoluzione di pensiero per la quale si sposta l'attenzione dal singolo, quale che esso sia[23], alle interazioni tra i singoli, le reti che questi formano, sta introducendo nell'intera Comunità Scientifica una nuova prospettiva con cui vedere la Scienza a tutto tondo[24].

Parimenti significativo, a nostro avviso, sempre nella formulazione di Hopfield delle reti neurali, è il concetto cardine di *diagramma di fase*, importato dalla Meccanica Statistica grazie agli sforzi di Amit, Gutfreund e Sompolinsky[25]: uno sguardo alla Figura 8 ci mostra che esiste una ben definita regione, nel piano rumore $T := 1/\beta$, carico della rete $\alpha = P/N$ (definito come il rapporto tra il numero di patterns che la rete deve gestire ed il numero di neuroni a disposizione), dove la rete funziona (cioè esegue spontaneamente e correttamente pattern recognition) chiamata *Retrieval Stable*, mentre fuori da quella regione la rete si comporta in maniera diversa: nella regione *Retrieval metastable* funziona con un certo margine di errore che dipendende significativamente dall'inizializzazione della rete; per livelli di rumore proibitivi (limite di alta T) la rete diventa un inutile paramagnete ergodico e per carichi troppo grandi (limite di alto α) diventa uno spin glass senza proprietà di richiamo. La conoscenza del diagramma di fase è di un'importanza cruciale per progettare una rete: per esempio, in questa formulazione con i patterns casuali, è inutile farle immagazzinare un numero di patterns uguale alla metà del numero di neuroni (cioè farla lavorare ad $\alpha = 0.5$) poiché, per quel valore di carico P/N, la rete non può riuscire ne ad imparare ne quindi a fare riconoscimento: è nostro credo

[23] si pensi nell'evoluzione della nostra cultura a quanto il *soggetto* sia stato centrale (e.g., il passaggio dall'antropocentrismo all'eliocentrismo).

[24] Si studiano ad oggi reti ovunque: reti sociali, reti di proteine, reti immunitarie, reti geniche, reti di telecomunicazioni e trasporti, etc. [19]

[25] Il lettore potrebbe questionare sull'impiego della misura di Gibbs (che sappiamo valere per la meccanica statistica canonica all'equilibrio) anche in un regime di stato stazionario fuori dall'equilibrio (si osservi che una configurazione di minimo del modello di Hopfield implichi, nel caso di pattern random in esame, che metà dei neuroni sia "on" e metà sia "off" e pertanto si verifichi un circolo di corrente stazionaria nella rete (un *treno di spikes*) fintanto che una $m^\mu \sim O(1)$). La prospettiva con cui Jaynes legge la meccanica statistica dovrebbe essere la chiave di lettura con cui vedere anche l'analisi meccanico statistica del modello di Hopfield (si veda a tal proposito la *lezione mancata* ed il contributo di Michele Castellana dedicato all'uso dell'inferenza di massima entropia).

che una conoscenza opportuna della meccanica statistica dei sistemi complessi possa offrire una chiave di lettura per governare il dilagare dell'IA nelle prossime decadi[26].

Volendo forzare la mano, ci sono due facetamente provocatorie riflessioni (che ben si prestano ad interpretare alcuni comportamenti bizzarri della nostra società): se si crede che queste reti Hebbiane possano effettivamente essere dei ragionevoli modelli di memoria associativa per i moduli corticali del cervello [11], il fatto che questi sistemi possano imparare qualunque cosa, compresi patterns composti unicamente da rumore bianco, fornisce uno spunto di riflessione. L'altro è che, se si aumenta l'esposizione di informazione alla rete (ci si sposta in α significativamente verso destra nel digramma di fase di Figura 8) la rete smette di funzionare opportunamente e confonde qualunque cosa, cosa che potrebbe fornire un'intuitiva spiegazione sull'apparente correlazione temporale tra l'avvento di internet e della globalizzazione (che ci hanno sovraesposti all'informazione) e la genesi dei complottisti...

Per chiudere, osserviamo che le reti neurali, in quanto particolari *reti elettriche*, sono state le prime reti biologiche ad essere state studiate a nostro avviso perchè nella prima metà del secolo scorso si aveva una conoscenza soddisfacente della Fisica classica (in particolare le equazioni di Maxwell ne sugellavano il trionfo e supplivano a perfezione le necessità della Fisiologia) mentre Biochimica e Genetica hanno avuto i loro primi trionfi nella seconda metà del secolo (e probabilmente molti ne avranno nel presente e nei prossimi): ad oggi una fetta della comunità scientifica (alla quale anche gli autori del presente scritto divulgativo appartengono [20]) è dedita allo scovare meccanismi Hebbiani, in ultima istanza forme di cognizione, anche in reti biochimiche (ed a diverse le scale): quando il sistema immunitario intraprende una specifica risposta non riconosce forse l'antigene di qualche patogeno che vuole usarci? E quando, dopo averlo eliminato, evita che ci re-infettiamo (proteggendoci con dei linfociti ad uopo) non sviluppa forse memoria?

Mentre le risposte *si* sono ovvie in entrambi i casi, rendersi conto che anche questi lo faccia in maniera Hebbiana [21] è sorprendente e suggerisce una sorta di universalità di processazione d'informazione, sulla quale c'è ancora molto lavoro ancora da fare e che potrebbe portare ad una nuova auspicabile intersezione tra l'Intelligenza Artificiale e la Biocomplessità: nella prossima medicina di precisione si sta tentando per certi aspetti una prima sintesi di questo connubio.

Infine, come ulteriore approfondimento rimandiamo con piacere al volume speciale che il *Journal of Physics A: mathematical & theoretical* ha dedicato quest'anno al tema *statistical physics & machine learning* [22].

[26] Si veda, a tal proposito, il contributo di Jean Barbier nel presente volume.

[1] M. Mezard, G. Parisi, M.A. Virasoro: *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, World Scientific Publishing Company, Singapore (1987).

[2] J.J. Hopfield: *Neural networks and physical systems with emergent collective computational abilities*, Proceedings of the National Academy of Sciences, 79 (1982) 2554.

[3] G. Parisi: *Infinite number of order parameters for spin-glasses*, Physical Review Letters 43 (1979) 1754.

[4] D.J. Amit, H. Gutfreund, H. Sompolinsky: *Storing infinite numbers of patterns in a spin-glass model of neural networks*, Physical Review Letters 55 (1985) 1530.

[5] D.J. Amit: *Modeling brain function: the world of attractor neural networks*, Cambridge University press, Cambridge (UK) (1992).

[6] A.C.C. Coolen, R. Kuhen, P. Sollich: *Theory of neural information processing systems*, Oxford University Press, Oxford (2005).

[7] L. Zdeborova: *Understanding deep learning is also a job for physicists*, 16 Nature Physics (2020) 1.

[8] H.S. Seung, H. Sompolinsky, N. Tishby: *Statistical mechanics of learning from examples*, Physical review A 45 (1992) 6056.

[9] Y. LeCun, Y. Bengio, G.E. Hinton: *Deep learning*, Nature 521 (2015) 436.

[10] D. Hebb: *The organization of behavior: a neuropsychological theory*, Psychology Press, Road Hove (UK) (2005).

[11] E. Schneidman, et al.: *Weak pairwise correlations imply strongly correlated network states in a neural population*, Nature 440 (2006) 1007.

[12] A. Barra, G. Genovese, P. Sollich, D. Tantari: *Phase transitions in Restricted Boltzmann Machines with generic priors*, Physical Review E 96 (2017) 042156.

[13] S. Kirkpatrick, M. Vecchi: *Optimization by simulated annealing*, Science 220 (1983) 671.

[14] D.H. Ackley, G.E. Hinton, T.J. Sejnowski: *A learning algorithm for Boltzmann machines*, Cognitive Science 9 (1985) 147.

[15] R. Salakhutdinov, G.E. Hinton, *Deep Boltzmann machines*, Proc. International Conference on Artificial Intelligence and Statistics (AISTATS) Clearwater Beach, Florida, USA. Volume 5 of JMLR:W&CP 5.(2009).

[16] A. Barra, A. Bernacchia, E. Santucci, P. Contucci: *On the equivalence of Hopfield networks and Boltzmann machines*, Neural Networks 34 (2012) 1.

[17] E. Agliari, A. Annibale, A. Barra, A.C.C. Coolen, T. Tantari: *Multitasking associative networks*, Physical Review Letters 109 (2012) 268101.

[18] M.E.J. Newman: *The structure and function of complex networks*, SIAM review 45 (2003) 167.

[19] G. Caldarelli: *Scale-free networks: complex webs in nature and technology*, Oxford University Press, Oxford (2007).

[20] E. Agliari, A. Barra, L. Dello Schiavo, A. Moro: *Complete integrability of information processing by biochemical reactions*, Scientific Reports 6 (2016) 1.

[21] E. Agliari, A. Annibale, A. Barra, A.C.C. Coolen, D. Tantari: *Retrieving infinite numbers of patterns in a spin-glass model of immune networks*, Europhysics Letters 117 (2017) 28003.

[22] E. Agliari, A. Barra, P. Sollich, L. Zdeborova (Editors): *Machine Learning and Statistical Physics: Theory, Inspiration, Application*, Journal of Physics A: Special Issue (2020).

Elena Agliari: è ricercatrice in Fisica Matematica presso Sapienza Università di Roma, dove insegna -tra i vari- *Modelli di Reti Neurali*. Si occupa principalmente di Meccanica Statistica dei Sistemi Complessi, Teoria dei Grafi e Processi Stocastici, con particolare attenzione alle loro applicazioni nella Biologia e nell'Intelligenza Artificiale.

Adriano Barra: è professore associato in Fisica Matematica presso l'Università del Salento, dove insegna -tra gli altri- *Metodi Matematici per l'Intelligenza Artificiale*. Si occupa di principalmente di Meccanica Statistica dei Sistemi Complessi, Teoria dei Grafi e Processi Stocastici, con particolare attenzione alle loro applicazioni nella Biologia e nell'Intelligenza Artificiale.

Reti neurali e forme di apprendimento

Il più bello dei mari è quello che non navigammo [...]
e quello che vorrei dirti di più bello non te l'ho ancora detto.
Nazim Hikmet

Daniele Tantari *Dipartimento di Matematica - Alma Mater Studiorum Bologna , Bologna, Italy*

In che modo il cervello umano apprende? Che caratteristiche deve avere un dispositivo in grado di apprendere come il cervello umano? A partire da queste due domande i modelli di rete neurale e le reti neurali artificiali stanno ripercorrendo lo stesso percorso formativo di un giovane studente: da reti che sanno imparare a memoria il più possibile e il più velocemente possibile si sta arrivando a modelli e dispositivi più complessi in grado di categorizzare, elaborare mappe concettuali, persino formare concetti e immagini nuove, in un processo che ricorda, senza pretesa di esserlo, la creatività.

L'eterna battaglia dell'insegnante è riuscire a motivare i suoi studenti verso livelli più alti di apprendimento convincendoli del fatto che imparare a memoria non sempre conviene: è una pratica troppo rigida, fragile e con evidenti limiti di scala. Rigida perché processa l'informazione ad un dettaglio estremo e senza l'esercizio critico di distinguere gli elementi più caratterizzanti da quelli meno rilevanti: tutto è estremamente importante e spesso si finisce che niente lo è. Rigidità porta fragilità: quando si impara una nozione a memoria si dà per scontato che questa sia corretta e non ci sono meccanismi di difesa contro un'eventuale scarsa affidabilità della fonte. Fragilità significa anche che se si dimentica qualcosa si dimentica tutto, dato che non ci sono informazioni a cui si riesce a dare priorità. I limiti di scala si riferiscono invece al fatto che la capacità di un qualsiasi dispositivo di memoria o del nostro cervello sono certo incredibili ma non illimitate, e l'apprendimento mnemonico non è quello più economico, quindi non quello più efficace per processare una grande mole di informazione. Nella maggior parte dei casi la capacità di sintetizzare o strutturare l'informazione è fondamentale.

Memorizzazione

Ci sono situazioni in cui tutti i dettagli hanno davvero la stessa importanza e la capacità di memorizzare è certamente indispensabile e in molti casi insostituibile: basti pensare all' importanza di ricordare (o di avere memorizzati in un dispositivo di memoria) un numero di telefono o il codice della propria carta, o di avere dei ricordi dettagliati delle esperienze vissute (anche se spesso ciò che rimane è solo un'idea astratta del ricordo, asciugata di molti dettagli). L' apprendimento mnemonico è il primo passo indispensabi-

le per lo sviluppo di forme di apprendimento più complete. Imparare a memoria non significa solo fissare dei concetti da qualche parte nella mente, significa sopratutto avere una strategia di recupero efficace. Alcuni sistemi fisici, per le proprietà della loro dinamica spontanea, possono essere considerati e usati come dispositivi di memoria, nel senso appena descritto. Si parla ad esempio della memoria di una molla o di un materiale a proposito della loro capacità di tornare in una configurazione di equilibrio. Nel contesto delle reti neurali si usa il termine *content-addressable memory* o memoria associativa. Consideriamo l'evoluzione temporale di un sistema fisico, il cui stato ad ogni istante t è descritto da un generico insieme di coordinate $\boldsymbol{\sigma}(t) = (\sigma_1(t), \dots, \sigma_n(t)) \in \Sigma$. La dinamica ha dei punti limite o attrattori, diciamo $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_p \in \Sigma$, se, partendo sufficientemente vicini ad uno di questi punti, cioè $\boldsymbol{\sigma}(0) = \boldsymbol{\xi}^\mu + \boldsymbol{\epsilon}$, il sistema evolverà in modo che $\lim_{t\to\infty} \boldsymbol{\sigma}(t) = \boldsymbol{\xi}^\mu$. Se questo succede possiamo interpretare i punti limite come vettori di informazione salvati in memoria e la dinamica come il processo di recupero/ riaffioramento / completamento del ricordo che si attua attraverso la correzione degli errori di un punto di partenza impreciso $\boldsymbol{\sigma}(0)$, in cui il ricordo manca di qualche dettaglio (per esempio la sensazione di avere il ricordo sulla punta della lingua) oppure è ancora molto confuso. Ogni sistema con dinamica ad attrattori è dunque un dispositivo che impara a memoria fintanto che abbiamo la possibilità di inserire vettori di informazione come punti limite della dinamica.

Reti neurali che imparano a memoria

Una rete neurale ricorrente è un sistema dinamico composto da n unità (o neuroni) interagenti $\sigma_1, \dots, \sigma_n$ introdotto come modello semplificato di una rete neuronale biologica diventando poi paradigma per i modelli di reti neurali artificiali. In accordo al modello discreto di McCulloch-Pitts [1], ogni neurone può trovarsi in due stati: $\sigma_i = 1$ quando il neurone emette segnale (*firing*) mentre $\sigma_i = -1$ se il neurone è a riposo. Lo stato del sistema è quindi definito al tempo t dal vettore di coordinate $\boldsymbol{\sigma}(t) = (\sigma_1(t), \dots, \sigma_n(t)) \in \Sigma = \{-1, 1\}^n$. Quando due neuroni i e j sono connessi da una sinapsi, si descrive la forza e il tipo di interazione con la variabile $J_{ij} \in \mathbb{R}$ e ogni neurone riceve un segnale dai neuroni a cui è connesso pari a $h_i = \sum_{j=1}^n J_{ij}\sigma_j$. Ogni neurone possiede una soglia $\theta_i \in \mathbb{R}$ e ad ogni tempo reagisce al segnale in ingresso ribilanciandosi, cioè emettendo segnale se quello in ingresso supera la soglia, $\sigma_i = +1$ se $h_i > \theta_i$, rimanendo a riposo altrimenti, $\sigma_i = -1$ se $h_i < \theta_i$. La sinapsi sarà quindi eccitatoria se $J_{ij} > 0$ o inibitoria se $J_{ij} < 0$ mentre due neuroni non connessi avranno $J_{ij} = 0$. Questo è un modello molto semplificato di una rete neurale biologica e tralascia molti dettagli microscopici riguardanti la struttura spaziale del neurone e delle sinapsi, l'elettrochimica dell'interazione, la presenza di ritardi nella trasmissione del segnale. L'effetto di tutti questi dettagli è usualmente modellizzato dalla presenza di stocasticità nella dinamica che provoca fluttuazioni dalla regola deterministica introdotta precedentemente, per cui

$$\sigma_i(t+1) = \text{sign}\left(h_i(\boldsymbol{\sigma}(t)) - \theta_i + \beta^{-1} z_i(t)\right), \quad (1)$$

dove $\text{sign}(x) = 1, 0, -1$ se x è positivo, nullo o negativo, rispettivamente, $z_i(t)$ sono variabili aleatorie indipendenti e β^{-1} descrive il livello di stocasticità/rumore della dinamica e viene denominata temperatura del sistema. Assumeremo da qui in avanti per brevità che $\theta_i = 0$. Se la distribuzione del rumore è simmetrica, $\mathbb{P}(z \in A) = \mathbb{P}(-z \in A)$, possiamo scrivere in termini probabilistici che

$$\mathbb{P}(\sigma_i(t+1)|\boldsymbol{\sigma}(t)) = F(\beta\sigma_i(t+1)h_i(t)) \quad (2)$$

dove $F(.)$ è la funzione di distribuzione del rumore, $F(z) = \mathbb{P}(z_i(t) < z)$. Una scelta comune è una funzione di distribuzione sigmoidale $F(z) = \frac{1}{2}[1 + \tanh(z)] = e^z/2\cosh(z)$. In una rete neurale ricorrente va inoltre specificato in quale ordine i neuroni si aggiornano. Esistono due casi estremi:

- aggiornamento sequenziale: ad ogni tempo viene scelto a caso un neurone e aggiornato,

$$\mathbb{P}(\sigma_i(t+1)) = \frac{1 + \sigma_i(t+1)\tanh(\beta h_i(t))}{2} \quad (3)$$

- aggiornamento parallelo: ad ogni tempo tutti i neuroni sono aggiornati contemporanea-

mente,

$$\mathbb{P}(\boldsymbol{\sigma}(t+1)) = \frac{e^{\beta \sum_{ij=1}^{n} J_{ij}\sigma_i(t+1)\sigma_j(t)}}{\prod_{i=1}^{n} 2\cosh(\beta h_i(t))} \quad (4)$$

Una rete neurale ricorrente si trova in letteratura con nomi diversi in contesti diversi: *Kinetic Ising Model* [2, 3, 7, 5, 6] in Meccanica Statistica o *Discrete Vector Autoregressive process* in Econometria [7] ed è utilizzata come strumento di analisi di serie temporali multivariate in diversi ambiti, da quello biologico [8, 9, 10, 11, 12] a quello economico e finanziario [13, 14, 15, 16, 17]. Le proprietà della dinamica di una rete ricorrente dipendono fortemente dagli accoppiamenti. Consideriamo per semplicità il caso di una dinamica deterministica, $\beta^{-1} = 0$, e guardiamo qualche esempio in dettaglio:

- $J_{ij} = 0$, $\theta_i \neq 0$: in questo caso i neuroni non interagiscono e il loro stato dipenderà esclusivamente dai valori delle loro soglie. Infatti $\sigma_i(t+1) = -\operatorname{sign}(\theta_i)$, per ogni i, ad ogni tempo $t > 0$ e qualunque sia lo stato iniziale del sistema. $-\boldsymbol{\theta}$ è l'attrattore (globale) della dinamica.

- $J_{ij} = J/n > 0$, $\theta_i = 0$: in questo caso le interazioni sono tutte di tipo eccitatorio e spingono i neuroni ad imitarsi a vicenda. La dinamica $\sigma_i(t+1) = \operatorname{sign}(\frac{1}{n}\sum_j \sigma_j(t)) = \operatorname{sign}(\frac{1}{n}\sum_j \sigma_j(0))$ tende naturalmente su due stati speciali, $\boldsymbol{\sigma}^+ = 1$ quando $m(0) = \frac{1}{n}\sum_i \sigma_i(0) > 0$ e $\boldsymbol{\sigma}^- = -1$ altrimenti, che sono i due punti limite.

- $J_{ij} = J/n < 0$, $\theta_i = 0$ e dinamica parallela: in questo caso $\sigma_i(t+1) = -\operatorname{sign}(\frac{1}{n}\sum_j \sigma_j(t)) = (-1)^t \operatorname{sign}(m(0))$ e la dinamica finisce in un ciclo limite di periodo 2 oscillando indefinitamente tra gli stati $\boldsymbol{\sigma}^+$ e $\boldsymbol{\sigma}^-$.

Più in generale si può dimostrare [19] che, se le interazioni sono simmetriche $J_{ij} = J_{ji}$, da qualsiasi punto iniziale il sistema converge su un punto limite se la dinamica è sequenziale e senza autointerazioni $J_{ii} = 0$ e al più su un ciclo se la dinamica è parallela. Quindi una rete ricorrente ha una dinamica ad attrattori e può essere usata come dispositivo di memoria se siamo in grado di controllare i punti limite della dinamica.

La regola di Hebb

Supponiamo di voler memorizzare i vettori di informazione

$$\boldsymbol{\xi}^\mu = (\xi_i^\mu, \dots, \xi_n^\mu) \in \Sigma, \quad \mu = 1, \dots, p. \quad (5)$$

Dobbiamo costruire una rete di interazioni J_{ij} (e soglie θ_i) in modo che si formino p punti limite nella dinamica della rete in corrispondenza dei vettori di informazione, con un bacino di attrazione non nullo affinché la rete sia in grado di recuperare i vettori memorizzati attraverso la sua dinamica spontanea. Il processo di costruzione delle interazioni ottimali è detto *learning*. Se invece pensiamo alla memorizzazione in un dispositivo artificiale, qualsiasi ricetta per la costruzione da zero delle interazioni che permetta di costruire i punti limite desiderati è una ricetta utile. Se pensiamo al *learning* come al processo biologico per cui la rete neuronale riceve nuovi stimoli (nozioni, esperienze) e le memorizza modificando le proprietà delle sue sinapsi, allora deve essere un processo costruttivo: a partire da una struttura sinaptica esistente J_{ij} (eredità, conoscenze acquisite in passato), la rete si modifica all'arrivo di un nuovo ricordo aggiornando le sue connessioni, ΔJ_{ij}. La regola di aggiornamento deve essere `locale`, cioè deve dipendere da informazioni accessibili alla sinapsi J_{ij}, quindi ai neuroni i e j. Un esempio di regola di aggiornamento sinaptico biologicamente fondata è la cosiddetta `Regola di Hebb`: una rete neurale ricorrente senza alcuna passata esperienza, $J_{ij} = 0$, impara a memoria un vettore di informazione $\boldsymbol{\xi} \in \Sigma$ aggiornando le sinapsi come

$$\Delta J_{ij} = \frac{\xi_i \xi_j}{n}. \quad (6)$$

In questo modo due neuroni che richiediamo essere nello stesso stato nel punto limite tenderanno ad imitarsi grazie ad un incremento positivo della loro sinapsi, mentre tenderanno ad antiimitarsi altrimenti. Il risultato è che la dinamica della rete

$$\sigma_i(t+1) = \operatorname{sign}(\xi_i)\operatorname{sign}(\frac{1}{n}\sum_j \xi_j \sigma_j(t)) = \quad (7)$$

porterà ogni stato iniziale su $\boldsymbol{\sigma} = \boldsymbol{\xi}$ se al tempo zero $m(0) = \frac{1}{n}\sum_j \xi_j \sigma_j(0) > 0$, cioè se lo stato

Mi sforzo ma non ricordo!

Non sempre riusciamo a far affiorare il ricordo se non abbiamo informazioni sufficienti a risvegliarlo. La presenza delle soglie θ_i permette di modellizzare questa situazione. Supponiamo che $J_{ij} = \xi_i\xi_j/n$, con $\sum_i \xi_i = 0$, e $\theta_i = \theta$, $|\theta| < 1$. Se definiamo $m(t) = \frac{1}{n}\sum_i \xi_i\sigma_i(t)$ avremo che

$$m(t+1) = \frac{1}{2}\,\mathrm{sign}(m(t)+\theta)+\frac{1}{2}\,\mathrm{sign}(m(t)-\theta)$$

cioè

$$m(t) = \begin{cases} 1 & \text{se } m(0) > |\theta| \\ -1 & \text{se } m(0) < -|\theta| \\ 0 & \text{se } |m(0)| < |\theta|, \end{cases} \qquad (8)$$

cioè se l'informazione iniziale non è sufficiente (più grande della soglia), la rete finisce su un punto limite corrispondente a nessun ricordo.

della rete $\boldsymbol{\sigma}(0)$ possiede un (seppur piccolo ma positivo) grado di sovrapposizione con il ricordo $\boldsymbol{\xi}$. Come effetto collaterale si è creato un altro stato limite $\boldsymbol{\sigma} = -\boldsymbol{\xi}$ a cui convergono gli stati con $m(0) < 0$. Generalizzando al caso di p vettori di informazione da imparare a memoria avremo che la rete creata avrà interazioni

$$J_{ij} = \frac{1}{n}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu. \qquad (9)$$

Nel caso in cui i vettori siano mutuamente ortogonali, cioè $\frac{1}{n}\boldsymbol{\xi}^\mu \cdot \boldsymbol{\xi}^\nu = \delta_{\mu\nu}$ che implica necessariamente $p < N$, la rete ricorrente è nota con il nome di Modello di Hopfield [20] e si dimostra che i p vettori di informazione sono tutti stati stazionari e attrattori della dinamica. Se lo stato iniziale possiede informazioni a sufficienza su uno dei vettori memorizzati, diciamo $\boldsymbol{\xi}^\mu$, cioè $m^\mu(0) = \sum_i \xi_i^\mu\sigma_i(0) > 0$, allora la rete riesce spontaneamente a ricordare e a ricostruire perfettamente il ricordo, come in Figura 1 (pannello intermedio).

Oltre agli stati $\pm\boldsymbol{\xi}^\mu$, `memorie fondamentali`, esistono anche altri attrattori della dinamica che corrispondono a misture dei p vettori di informazione fondamentali. Nel contesto della memorizzazione corrispondono a ricordi che si intrecciano, nozioni che si confondono, immagini che si accavallano e sono spesso considerate fallimenti dell'apprendimento, vedi Figura 1 (pannello inferiore). Vedremo nella prossima sezione che la creazione di immagini nuove è invece la capacità fondamentale di una rete neurale che apprende.

Se le memorie fondamentali non sono esattamente ortogonali (questo succede se i ricordi sono variabili aleatorie indipendenti o a maggior ragione se sono correlati), la situazione si complica: non solo nascono nuovi attrattori ma allo stesso tempo le memorie fondamentali possono destabilizzarsi e la rete non è più in grado di ricostruirle esattamente. Il fenomeno è detto *crosstalk*, termine che si riferisce al segnale di disturbo su una memoria dovuto alle interferenze delle altre. Abbiamo infatti che una memoria $\boldsymbol{\xi}^\mu$ è un punto fisso della dinamica solo se $\xi_i^\mu = \mathrm{sign}(\sum_j J_{ij}\xi_j^\mu)$, dove

$$(\boldsymbol{J}\boldsymbol{\xi}^\mu)_i = \xi_i^\mu + \sum_{\nu\neq\mu}^{p} \xi_i^\nu \frac{\boldsymbol{\xi}^\mu \cdot \boldsymbol{\xi}^\nu}{n}. \qquad (10)$$

Il secondo dei termini a destra rappresenta il contributo di *crosstalk* ed è nullo solo se i vettori sono ortogonali. In generale è tanto più grande quanto più i vettori sono correlati e anche nel caso di vettori *random* indipendenti è un contributo a media nulla ma con una varianza che cresce al crescere del numero di vettori/ricordi p che stiamo cercando di memorizzare, il cosiddetto `carico` della rete. Si definisce quindi la `capacità` il massimo numero di vettori di informazioni che la rete è in grado di imparare a memoria e quindi di ricostruire spontaneamente. Nel caso di vettori random scorrelati e a media nulla la capacità del modello di Hopfield cresce come $n/4\log n$ [21]. Nel caso in cui ci accontentassimo di una ricostruzione non perfetta, che preveda qualche errore qui e lì, allora la capacità cresce come una costante (piccola) per n.

La prospettiva meccanico-statistica

Questo è quello che succede a maggior ragione quando consideriamo una dinamica stocastica, con $\beta^{-1} > 0$. In questo caso la dinamica non può avere punti fissi in quanto c'è sempre una probabilità non nulla che qualche neurone cambi il suo stato. C'è bisogno di un linguaggio diver-

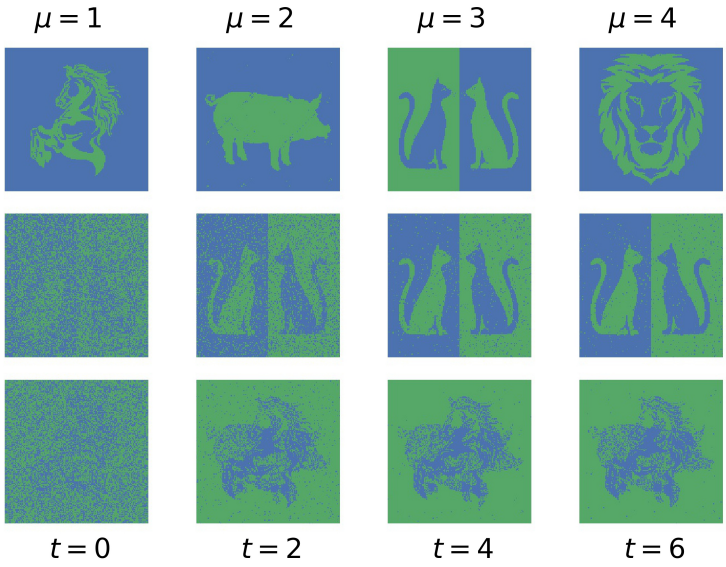


Figura 1: *Dinamica di richiamo in un Modello di Hopfield con $p = 4$ vettori di informazione (immagini) memorizzate (pannello superiore). A partire da uno stato iniziale rumoroso la rete riesce a richiamare alla memoria una delle immagini (pannello intermedio). A causa della correlazione tra le immagini memorizzate a volte la rete finisce su uno stato spurio (pannello inferiore).*

so per descrivere le proprietà della rete, che è quello della probabilità e più in particolare della Meccanica Statistica. L'equazione (3) ad esempio definisce la probabilità di transizione di una catena di Markov che ammette una distribuzione d'equilibrio

$$P_\beta(\boldsymbol{\sigma}) = Z^{-1} \exp\left(\beta \sum_{i<j=1}^{n} J_{ij}\sigma_i\sigma_j\right) \qquad (11)$$

su Σ, dove $Z = \sum_\sigma \exp\left(\beta \sum_{ij}^n J_{ij}\sigma_i\sigma_j\right)$, costante di normalizzazione che assicura $\sum_\sigma P_\beta(\sigma) = 1$, è detta funzione di partizione. Questo significa che le traiettorie del sistema, che descrivono lo stato della rete nel tempo, esplorano lo spazio di tutti i possibili ricordi Σ, ricostruendone ognuno con probabilità $P_\beta(\boldsymbol{\sigma})$. Quest'ultima è detta distribuzione di Boltzmann-Gibbs e conferisce alla rete ricorrente anche il titolo di `Macchina di Boltzmann`, che si aggiunge ai precedenti e viene usato spesso nel contesto delle reti neurali artificiali. Non possiamo più studiare i punti limite di una dinamica deterministica, però ha senso individuare le regioni visitate quasi certamente o con probabilità molto alta. Nel caso del modello di Hopfield, in cui $J_{ij} = \frac{1}{n}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu$, quando il rumore stocastico non è troppo alto ($\beta \gg 1$) e il carico della rete è sufficientemente basso $p \ll n$, si dimostra che le configurazioni visitate tipicamente formano dei cluster, ed esistono cluster centrati sui vettori di informazione $\{\boldsymbol{\xi}^\mu\}$, cioè in cui

$$\langle m^\mu(\boldsymbol{\sigma})\rangle = \left\langle \frac{1}{n}\sum_{i=1}^{n} \xi_i^\mu \sigma_i \right\rangle = m_\beta^\mu > 0 \qquad (12)$$

e $m_\beta^\nu \sim 0$ per $\nu \neq \mu$. Questi cluster sono detti stati puri e il regime appena descritto viene detto di `richiamo`, in quanto la rete sta imparando a memoria qualitativamente come nel caso deterministico, richiamando alla mente i vettori di informazione immagazzinati nella struttura sinaptica e in modo spontaneo, infatti per $t \gg 1$

$$m^\mu(t) = \frac{1}{n}\sum_{i=1}^{n} \xi_i^\mu \sigma_i(t) \sim m_b^\mu > 0, \qquad (13)$$

e $m^\nu(t) \sim 0$ per $\nu \neq \mu$. Tuttavia se la temperatura della dinamica o il carico della rete diventano troppo alti, nascono due regimi di lavoro differenti:

- regime `paramagnetico`: in questo regime la temperatura della rete è troppo alta ed ogni neurone si comporta perlopiù in modo casuale senza bloccarsi in configurazioni specifiche. Avremo di conseguenza che per ogni μ, $\langle m^\mu(\boldsymbol{\sigma})\rangle \sim 0$, cioè la rete non riesce a richiamare nessun ricordo ma fluttua indefinitamente.

- regime vetroso o di `spin glass`: in questo regime il carico della rete e quindi le interferenze tra i ricordi sono eccessivi. I neuroni si bloccano su configurazioni specifiche che

però sono solo debolmente correlate con i ricordi, cioè ancora una volta, ma per motivi diversi $\langle m^\mu(\boldsymbol{\sigma})\rangle \sim 0$ per ogni μ. In termini di cluster, gli stati puri spariscono così come ogni altro gruppo di configurazioni tipiche correlate con i ricordi.

La capacità del modello di Hopfield come dispositivo che impara a memoria è riassunto quindi nel diagramma di fase approssimato di Figura 2.

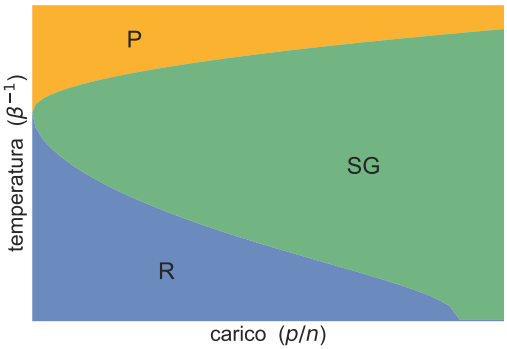


Figura 2: *Diagramma di fase del Modello di Hopfield nel piano carico-temperatura con i tre regimi: R) fase di richiamo, P) fase paramagnetica, SG) fase di spin glass.*

Apprendimento

Nell'apprendimento di alto livello l'obiettivo passa dalla memorizzazione/riproduzione di vettori di informazione alla ricerca di una struttura, un filo comune, di regole generali che li spieghino.

Se nell'apprendimento mnemonico i vettori di informazione (immagini, nozioni, etc.) sono dei modelli ideali da incamerare esattamente, intorno ai quali costruire la struttura sinaptica e a cui la rete/il dispositivo si riferisce ogni volta che viene stimolata differentemente (con la dinamica di richiamo dal punto iniziale al punto limite), nell'apprendimento vero e proprio diventano degli `esempi` di realtà, da studiare, strutturare, per elaborare un modello interno che la realtà la spieghi.

Se nell'apprendimento mnemonico il contesto ideale è rappresentato da pochi vettori di informazione da memorizzare, possibilmente scorrelati per evitare stati spuri, nell' apprendimento propriamente detto la cosa migliore è avere tanti esempi (con la loro struttura di correlazione) dai quali evincere un modello interno idealizzato.

Supponiamo di dover imparare a distinguere animali diversi. Una rete che impara a memoria ha bisogno, quasi biblicamente, di un'immagine ideale per specie da fissare alla mente: ogni immagine nuova richiamerà alla memoria una delle immagini modello, la più simile. Una rete che apprende invece osserva tantissime immagini di animali e le organizza imparando a riconoscerne la struttura di similarità e creando internamente prototipi o caratteristiche idealizzate. In questo senso si parla di apprendimento non supervisionato.

Macchine di Boltzmann che apprendono

La ricerca dei principi di apprendimento non supervisionato (biologico e artificiale) ha motivato lo sviluppo di una prospettiva differente sull'uso delle Macchine di Boltzmann [22, 23]. La distribuzione (11) assegna probabilità alta alle configurazioni $\boldsymbol{\sigma}$ che soddisfano il maggior numero di vincoli della forma $\sigma_i\sigma_j = \mathrm{sign}(J_{ij})$, cioè con energia

$$E = -\sum_{i<j}^{n} J_{ij}\sigma_i\sigma_j \tag{14}$$

più bassa. È la distribuzione di equilibrio di una dinamica stocastica in cui ogni neurone/unità si accende con probabilità (3), che può essere riscritta come

$$\mathbb{P}(\sigma_i(t+1) = 1) = \frac{1}{1 + e^{-\beta\Delta E_i(t)}}, \tag{15}$$

dove $\Delta E_i(t)$ è la differenza di energia (o di vincoli soddisfatti) se il neurone i-esimo passa da spento a acceso. In questo senso gli accoppiamenti possono essere interpretati come i vincoli che devono essere soddisfatti dalle configurazioni più visitate nel tempo dalla rete. Quando la rete riceve un vettore di informazione $\boldsymbol{\xi}^\mu$ i neuroni sono forzati (*clamped*) ad essere nella configurazione $\boldsymbol{\sigma} = \boldsymbol{\xi}^\mu$ e le sinapsi si muovono di conseguenza in un processo di learning. A differenza dell'apprendimento mnemonico in cui $\Delta J_{ij} \propto \xi_i^\mu\xi_j^\mu$, cioè ad ogni passo la rete si concentra per soddisfare i vincoli di un solo vettore di informazione alla volta, nell' apprendimento non supervisiona-

to la rete cerca di costruire una visione d'insieme, un modello interno per spiegare contemporaneamente tutti i vettori di informazione, che in questo contesto chiameremo esempi e il loro insieme *training set* $\mathcal{M} = \{\boldsymbol{\sigma}^a\}_{a=1}^m$[1]. Invece di chiedere rigidamente che i vettori di informazione siano punti limite della dinamica, chiederemo che le proprietà del training set siano riprodotte bene dalla rete, o in altre parole che la distribuzione di probabilità (empirica) dei vettori del training set sia ben approssimata da quella d'equilibrio della rete/macchina. Se $P_\beta(\boldsymbol{\sigma})$ è la distribuzione della macchina di Boltzmann e $\hat{P}(\boldsymbol{\sigma}) = \frac{1}{m}\sum_{a=1}^m \delta_{\boldsymbol{\sigma}\boldsymbol{\sigma}^a}$ la distribuzione empirica del training set, si può definire la distanza tra le due come

$$\mathcal{D}(\hat{P}||P_\beta) = \sum_{\boldsymbol{\sigma}} \hat{P}(\boldsymbol{\sigma}) \log \frac{\hat{P}(\boldsymbol{\sigma})}{P_\beta(\boldsymbol{\sigma})}. \qquad (16)$$

L'obiettivo di un apprendimento non supervisionato è minimizzare questa distanza, o equivalentemente, massimizzare la funzione

$$\mathcal{L} = \sum_{a=1}^m \log P_\beta(\boldsymbol{\sigma}^a), \qquad (17)$$

detta *log-likelihood*, che (logaritmo a parte) rappresenta la probabilità che tutto il training set di esempi sia generato e quindi spiegato dalla rete. Si può massimizzare iterativamente questa quantità calcolando la sua derivata rispetto agli accoppiamenti della rete

$$\frac{\partial \log P_\beta(\boldsymbol{\sigma}^a)}{\partial J_{ij}} = \beta\left(\sigma_i^a\sigma_j^a - \langle\sigma_i\sigma_j\rangle\right), \qquad (18)$$

dove $\langle . \rangle$ indica la media rispetto a P_β, e incrementando il valore delle sinapsi proporzionalmente

$$\Delta J_{ij} \propto \left(\sigma_i^a\sigma_j^a - \langle\sigma_i\sigma_j\rangle\right). \qquad (19)$$

Il metodo è detto *gradient ascent* perché la rete segue la direzione di massima crescita fornita dal gradiente ed è una generalizzazione della regola di Hebb (6). Come la regola di Hebb ha una forma locale e contiene il termine $\sigma_i^a\sigma_j^a$ (equivalente a $\xi_i^\mu\xi_j^\mu$). A differenza di essa non si esaurisce in m (equivalentemente p) passi ma si rinforza fino a convergenza, momento in cui $\langle\sigma_i\sigma_j\rangle$ sarà vicino a $\frac{1}{m}\sum_a \sigma_i^a\sigma_j^a$, cioè la rete avrà imparato la struttura di correlazione del training set. Se vogliamo una rete che impari delle correlazioni di ordine maggiore possiamo generalizzare la macchina di Boltzmann vista finora dividendo i neuroni in due gruppi con funzioni diverse: neuroni `visibili` $\boldsymbol{\sigma}$ e neuroni `nascosti` $\boldsymbol{\tau}$. I neuroni visibili sono l'interfaccia con l'ambiente esterno, quelli che vengono stimolati direttamente (e clamped) da un input di informazione: nelle reti viste finora tutti i neuroni erano visibili. I neuroni nascosti, *hidden units*, non interagiscono mai direttamente con l'ambiente esterno e servono alla rete per spiegare dei vincoli che non possono essere rappresentati da interazioni di coppia tra neuroni visibili, quindi per imparare una realtà più complessa. L'obiettivo della macchina rimane quello di riprodurre al meglio la distribuzione di probabilità empirica del training set $\hat{P}(\boldsymbol{\sigma})$ con la distribuzione d'equilibrio dei suoi neuroni visibili

$$P_\beta(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\tau}} P_\beta(\boldsymbol{\sigma}, \boldsymbol{\tau}), \qquad (20)$$

ottenuta marginalizzando su tutte le possibili configurazioni dei neuroni nascosti. Minimizzando di nuovo la distanza $\mathcal{D}(\hat{P}||P_\beta)$, o equivalentemente massimizzando la log-likelihood, si ottiene una regola di aggiornamento delle sinapsi analoga alla (18) essendo

$$\frac{\partial \log P_\beta(\boldsymbol{\sigma}^a)}{\partial J_{ij}} = \beta\left(\sigma_i^a\sigma_j^a - \langle\sigma_i\sigma_j\rangle\right), \qquad (21)$$

se J_{ij} connette due neuroni visibili. Se la sinapsi $J_{\mu\nu}$ connette due neuroni nascosti

$$\frac{\partial \log P_\beta(\boldsymbol{\sigma}^a)}{\partial J_{\mu\nu}} = \beta\left(\langle\tau_\mu\tau_\nu\rangle_a - \langle\tau_\mu\tau_\nu\rangle\right), \qquad (22)$$

dove $\langle . \rangle_a$ identifica la media rispetto a $P_\beta(\boldsymbol{\tau}|\boldsymbol{\sigma}^a)$, cioè la distribuzione d'equilibrio dei neuroni nascosti ottenuta condizionando i neuroni visibili (clamped) sull'input $\boldsymbol{\sigma}^a$. Infine se la sinapsi $J_{i\mu}$ connette un neurone visibile ed uno nascosto si avrà

$$\frac{\partial \log P_\beta(\boldsymbol{\sigma}^a)}{\partial J_{i\nu}} = \beta\left(\sigma_i^a\,\langle\tau_\nu\rangle_a - \langle\sigma_i\tau_\nu\rangle\right). \qquad (23)$$

[1] Cambiamo notazione, da $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^p$ a $\{\boldsymbol{\sigma}^a\}_{a=1}^m$ per sottolineare che la strategia di apprendimento sarà diversa e differenziare i vettori di informazione di un apprendimento mnemonico dagli esempi di un apprendomento non supervisionato, dove per altro tipicamente $m \gg p$

Si dimostra che la dinamica di apprendimento è convessa fintanto che ci sono solo neuroni visibili mentre possono nascere minimi locali quando la rete usa anche dei neuroni nascosti. Questi minimi locali corrispondono a diversi modi in cui la rete può usare le unità nascoste per spiegare i vincoli impliciti presenti nel training set. Alla fine dell'allenamento, proprio le sinapsi che collegano i neuroni nascosti a quelli visibili ci rivelano quello che la rete ha imparato e in che modo è stato codificato.

Macchine di Boltzmann ristrette

Per capire meglio consideriamo un esempio di rete con neuroni nascosti con una struttura semplice ma allo stesso tempo molto rilevante perché alla base delle moderne strutture di reti neurali artificiali: le Macchine di Boltzmann ristrette (RBM). In questo tipo di rete neurale le sinapsi collegano solamente neuroni visibili a neuroni nascosti, non essendo invece presenti interazioni dirette tra coppie di neuroni dello stesso tipo. Una RBM può quindi essere rappresentata da un grafo bipartito come in Figura 5 e Figura 6 e la sua distribuzione di equilibrio sarà

$$P_\beta(\boldsymbol{\sigma},\boldsymbol{\tau}) = Z_\beta^{-1} P_0(\boldsymbol{\tau}) \exp\left(\beta \sum_{i\mu} J_{i\mu}\sigma_i\tau^\mu\right)$$

dove abbiamo aggiunto il termine $P_0(\boldsymbol{\tau})$ che rappresenta la distribuzione a priori dei neuroni nascosti. La forma specifica di una RBM introduce numerosi vantaggi da un punto di vista computazionale nel calcolo del gradiente e quindi dell'aggiornamento delle sinapsi (23), derivanti principalmente dal fatto che i neuroni nascosti sono indipendenti condizionatamente allo stato dei visibili e viceversa. Da un punto di vista teorico la proprietà più interessante invece è che possiamo facilmente calcolare la distribuzione d'equilibrio sui neuroni visibili marginalizzando sui neuroni nascosti ottenendo

$$\begin{aligned} P_\beta^{\mathrm{RBM}}(\boldsymbol{\sigma}) &= Z_\beta^{-1}\mathbb{E}_{\boldsymbol{\tau}} \exp\left(\beta \sum_{i\mu} J_{i\mu}\sigma_i\tau^\mu\right) \\ &= Z_\beta^{-1} \exp\left(\sum_\mu \psi(\beta \sum_i J_{i\mu}\sigma_i)\right), \end{aligned} \tag{24}$$

dove $\psi(x) = \log \mathbb{E}_\tau[e^{\tau x}]$. La forma specifica dipende quindi dalla forma del *prior*, quindi dalla natura di neuroni nascosti. Supponiamo di avere neuroni nascosti gaussiani $\mathcal{N}(0,1)$, i.e. $P_0(\tau) = \frac{1}{\sqrt{2\pi}} e^{-\tau^2/2}$, allora $\psi(x) = x^2/2$ e

$$P_\beta^{\mathrm{RBM}}(\boldsymbol{\sigma}) = Z_\beta^{-1} \exp\left(\beta \sum_{i<j} \hat{J}_{ij}\sigma_i\sigma_j\right), \tag{25}$$

dove

$$\hat{J}_{ij} \propto \sum_\mu J_i^\mu J_j^\mu. \tag{26}$$

cioè una RBM ha la distribuzione d'equilibrio di un modello di Hopfield con vettori di informazione $\{\boldsymbol{J}^\mu\}$ [24, 25, 27] . Questo è il momento migliore quindi per confrontare un apprendimento non supervisionato (tramite RBM) con uno mnemonico (con apprendimento Hebbiano). A differenza dell'apprendimento mnemonico, in cui le sinapsi sono costituite direttamente dai vettori di infromazione/ modelli osservati $\{\boldsymbol{\xi}^\mu\}$ (disposti con la regola di Hebb), nell'apprendimento non supervisionato le componenti principali delle sinapsi sono i vettori $\{\boldsymbol{J}^\mu\}$ che dipendono dagli esempi $\{\boldsymbol{\sigma}^a\}$ solo tramite il processo di apprendimento (23). Si può dire che nell'apprendimento non supervisionato la rete impara i modelli ($\{\boldsymbol{J}^\mu\}$) dagli esempi invece che imparare a memoria gli esempi stessi. Finché gli esempi sono pochi e ortogonali allora i due apprendimenti sono tutto sommato equivalenti, ma quando gli esempi sono tanti e correlati tra loro la Macchina di Boltzmann imparerà i modelli $\{\boldsymbol{J}^\mu\}$ che sintetizzano al meglio l'informazione del *training set*.

Un esperimento controllato

Per capire meglio le capacità di apprendimento di una RBM possiamo eseguire un esperimento di tipo *teacher-student* [28, 29]. Un insegnante possiede l'informazione $\boldsymbol{\xi}$ che vuole trasmettere al suo studente, tuttavia non vuole che la impari a memoria bensì lo stimola ad apprendere con le sue forze concedendogli solo degli esempi su cui studiare. Un esempio $\boldsymbol{\sigma}^a$ è un'informazione parziale, o un po' distorta, con qualche dettaglio differente dalla nozione esatta costituita dal vettore $\boldsymbol{\xi}$, e può essere costruito a partire da esso semplicemente scambiando qualche bit con pro-

babilità $\epsilon \in (0, 1/2)$[2]: al crescere di ϵ l'esempio sarà via via più lontano dal suo modello mentre se ϵ è prossima a zero gli esempi saranno molto calzanti e porteranno molta più informazione sulla vera nozione da imparare $\boldsymbol{\xi}$. Lo studente riceve quindi un insieme di m esempi $\mathcal{M} = \{\boldsymbol{\sigma}^a\}_{a=1}^m$, il *training set*, da osservare e da cui apprendere. Supponiamo che il metodo di studio dello studente (comprensivo dei suoi dispositivi e il suo cervello) sia quello di una RBM con un neurone nascosto che apprende con lo schema di apprendimento non supervisionato introdotto precedentemente. Al termine dell'apprendimento le sue sinapsi saranno il vettore $\boldsymbol{J}$. Per capire l'efficacia del suo apprendimento dobbiamo valutare la distanza tra $\boldsymbol{J}$ e $\boldsymbol{\xi}$, l'informazione che lo studente non conosceva, e lo facciamo tramite la quantità

$$q = \langle \boldsymbol{J}, \boldsymbol{\xi} \rangle := \frac{1}{n} \sum_{i=1}^{n} \text{sign}(J_i)\xi_i : \qquad (27)$$

più alto è q migliore sarà stata la performance dello studente. In Figura 3 si può seguire la performance dello studente durante l'allenamento.

Quando la dimensione del vettore d'informazione è molto alta ($n \to \infty$) è molto difficile che lo studente riesca solo per caso ad imparare qualcosa, in questo caso si avrà sempre $q \sim 0$, quindi possiamo giudicare la performance positiva non appena $q > 0$. Possiamo distinguere tre regimi (Figura 4):

- $\epsilon \sim 0$: gli esempi $\boldsymbol{\sigma}^a$ saranno molto simili al modello $\boldsymbol{\xi}$ e lo studente può raggiungere un risultato positivo anche imparando a memoria uno degli esempi. In questo caso lo studente può prendere alla lettera ogni parola dell'insegnante, che è molto affidabile o molto (troppo?) buono. Questo è il caso in cui apprendimento mnemonico e non supervisionato sono parimenti efficaci.

- $\epsilon \sim 1/2$ e $m/n \ll 1$: gli esempi sono pochi e hanno pochissima informazione sul modello. In questo caso l'insegnante è un cattivo insegnante e né l'apprendimento mnemonico né quello non-supervisionato possono dare risultati positivi: semplicemente lo studente, anche se molto in gamba, non ha ricevuto l'informazione sufficiente per poter apprendere.

- $\epsilon \sim 1/2$ e $m/n \gg 1$: gli esempi hanno poca informazione sul modello ma sono molto numerosi. Questo è molto probabilmente il

[2] $\epsilon = 1/2$ corrisponde alla perdita totale dell'informazione, se ϵ fosse maggiore gli esempi risultanti sarebbero tipicamente dei negativi, che portano la stessa informazione dell'originale

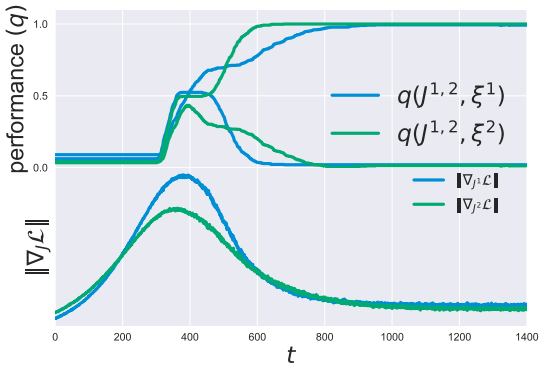


Figura 3: *Monitoraggio della fase di allenamento di una RBM in un esperimento teacher-student. La rete ha 2 neuroni nascosti e deve apprendere 2 vettori di informazione ortogonali tra loro. Quando il gradiente si stabilizza su un minimo le sinapsi allenate $\boldsymbol{J}^1$ e $\boldsymbol{J}^2$ hanno imparato i due vettori di informazione dell' insegnante $\boldsymbol{\xi}^1$ e $\boldsymbol{\xi}^2$*

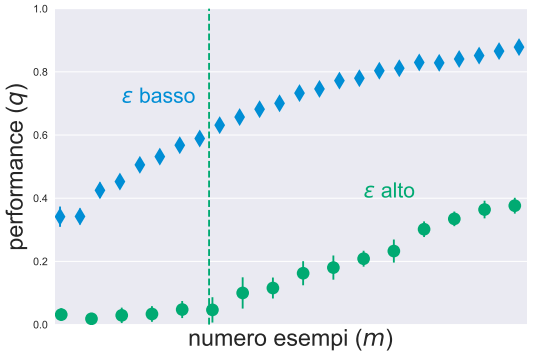


Figura 4: *Esperimento teacher-student e performance dello studente al variare del numero di esempi a disposizione m. Lo studente apprende allenando una RBM su un training set di esempi con rumore ϵ. Finchè il rumore è basso lo studente apprende con pochi esempi e l' apprendimento è mnemonico. Quando il rumore è alto l' apprendimento mnemonico non è efficace e lo studente necessita di un numero di esempi superiori ad una certa soglia.*

caso di un insegnante che vuole stimolare lo studente, che non dà l'informazione tutta insieme ma distribuita su un numero sufficientemente alto di esempi. Se lo studente imparasse a memoria gli esempi ($\boldsymbol{J} = \boldsymbol{\sigma}^a$), non otterrebbe un risultato positivo, perché $q = \langle \boldsymbol{J}, \boldsymbol{\xi} \rangle = \langle \boldsymbol{\sigma}^a, \boldsymbol{\xi} \rangle \sim 1 - 2\epsilon \sim 0$. Il risultato dell'apprendimento tramite RBM è invece positivo.

Le stesse considerazioni si generalizzano al caso di una coppia *teacher-student* con p vettori di informazioni e p neuroni nascosti [28, 29].

Prototipi e caratteristiche

L' esperimento precedente rappresenta un caso molto speciale di apprendimento in cui ciò che si deve imparare è uno o più prototipi ideali e il numero di neuroni nascosti della rete è proprio uguale al numero di prototipi da imparare. Ci sono allora due classi di domande ulteriori che ci si può porre:

- La ricerca di prototipi ideali è l'unica forma di apprendimento?
- Come la struttura della rete (per esempio il numero di neuroni nascosti) influenza l' apprendimento, cioè il modello interno con cui la rete interpreta e spiega la realtà ?

Riguardo la prima questione esistono almeno due modalità differenti di apprendimento o *pattern recognition*: la formazione di `prototipi` e la combinazione di `caratteristiche`. Nel primo caso, come nell'esempio precedente, la rete impara a riconoscere gli oggetti nel complesso, costruendo dei modelli ideali, appunto i prototipi, interpretando la realtà come realizzazione imperfetta di essi. Nel secondo caso la rete impara a riconoscere, estraendole dagli esempi e codificandole nelle sinapsi, delle caratteristiche ed interpreta la realtà come combinazione di esse. Ad esempio una rete può imparare a distinguere i gatti dai maiali costruendo due immagini interne ideali di felinità e suinità (Figura 5) oppure combinando insieme immagini interne di caratteristiche come la forma della coda, del muso etc... (Figura 6)

Una RBM che ha appreso dei prototipi è un modello di Hopfield che genera cluster di configurazioni centrate su $\boldsymbol{J}^\mu$ (i prototipi). In ogni cluster,

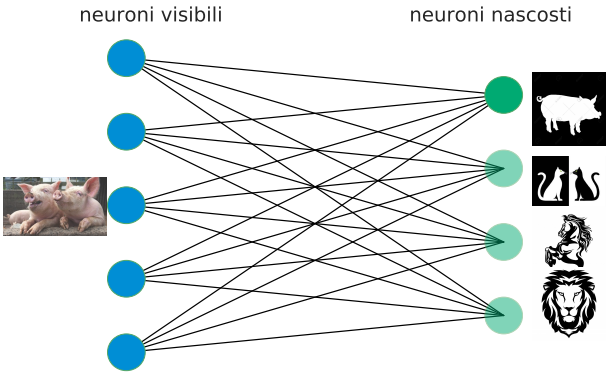


Figura 5: *Schema di una RBM che apprende per prototipi: l' immagine sui neuroni visibili è interpretata come variante di uno dei prototipi memorizzati. Il neurone nascosto corrispondente è acceso mentre tutti gli altri sono quasi spenti.*

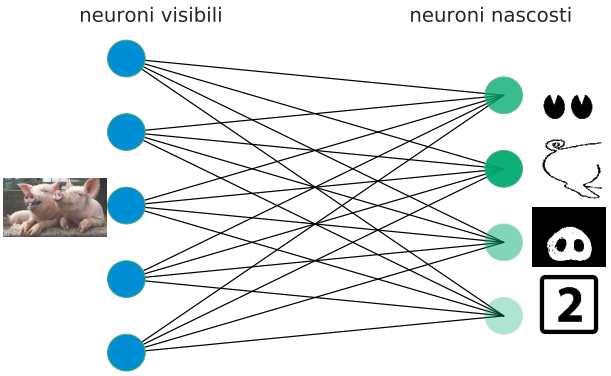


Figura 6: *Schema di una RBM che apprende per caratteristiche: l' immagine sui neuroni visibili è decomposta e interpretata come sovrapposizione delle caratteristiche memorizzate. Più neuroni nascosti, quelli corrispondenti alle caratteristiche utilizzate, sono accesi.*

diciamo il μ- esimo, un solo neurone nascosto τ^μ è acceso, stimolato da

$$P(\tau^\mu|\boldsymbol{\sigma}) = Z_\mu^{-1} \exp\left(\beta m^\mu(\boldsymbol{\sigma})\tau^\mu\right), \qquad (28)$$

dove $m^\mu(\boldsymbol{\sigma}) = \boldsymbol{J}^\mu \cdot \boldsymbol{\sigma} \gg 0$, mentre tutti gli altri sono spenti essendo $m^\nu(\boldsymbol{\sigma}) \ll 1$ per $\nu \neq \mu$. In una RBM che ha estratto caratteristiche i clusters sono centrati su stati spuri che sono combinazioni dei vettori $\boldsymbol{J}^\mu$, questa volta interpretabili come caratteristiche. In questo caso in ogni cluster più di un neurone nascosto è acceso, anche se più debolmente, essendo $m^\mu(\boldsymbol{\sigma}) > 0$ per ogni μ corrispondente ad una caratteristica $\boldsymbol{J}^\mu$ presente nel cluster.

Riguardo la seconda questione, sicuramente la tipologia di apprendimento è influenzata fortemente dai parametri che regolano l'architettura della rete come il numero di neuroni nascosti, la loro natura e la presenza di vincoli sulla struttura sinaptica. In particolare si può mostrare come al variare di questi parametri una RBM allenata su uno stesso training set di esempi passi da un apprendimento per prototipi ad uno per caratteristiche [30, 31].

Va sottolineato infine che un apprendimento di alto livello deve necessariamente saper integrare entrambe le modalità, estraendo caratteristiche e poi combinandole per creare caratteristiche di livello sempre più alto fino a costruire dei prototipi ideali. Per questo motivo i dispositivi più efficienti di reti neurali che apprendono percorrono il paradigma del *Deep Learning*, in cui più RBM sono impilate, e allenate, una dopo l'altra [32].

Abbiamo visto reti neurali che imparano a memoria immagini, abbiamo visto reti neurali che apprendono tramite immagini. Memorizzazione come registrazione, apprendimento come rielaborazione. Ma per una teoria dell'immaginazione propriamente detta manca ancora tanta fantasia.

[1] W. S. McCulloch, W. Pitts: *A logical calculus of the ideas immanent in nervous activity*, The bulletin of mathematical biophysics, 5 (1943) 115.

[2] B. Derrida, E. Gardner, A. Zippelius: *An exactly solvable asymmetric neural network model*, EPL (Europhysics Letters), 4 (1987) 167.

[3] A. Crisanti, H. Sompolinsky: *Dynamics of spin systems with randomly asymmetric bonds: Ising spins and Glauber dynamics*, Physical Review A, 37 (1988) 4865.

[4] B. Dunn,Y. Roudi: *Learning and inference in a nonequilibrium ising model with hidden nodes.* Physical Review E, 87 (2013) 022127.

[5] A. Decelle, Pan Zhang: *Inference of the sparse kinetic ising model using the decimation method*, Physical Review E, 91 (2015) 052136.

[6] C. Campajola, F. Lillo, D. Tantari: *Inference of the kinetic ising model with heterogeneous missing data*, Physical Review E, 99 (2019) 062138.

[7] C. Campajola, F. Lillo, P. Mazzarisi, D. Tantari, *On the equivalence between the Kinetic Ising Model and discrete autoregressive processes*, arXiv preprint arXiv:2008.10666 (2020)

[8] S. Cocco, R. Monasson, L. Posani, G. Tavoni: *Functional networks from inverse modeling of neural population activity*, Current Opinion in Systems Biology, 3 (2017) 103.

[9] T.-A. Nghiem, B. Telenczuk, O. Marre, A. Destexhe, U. Ferrari: *Maximum-entropy models reveal the excitatory and inhibitory correlation structures in cortical neuronal activity*, Physical Review E, 98 (2018) 012402.

[10] U. Ferrari, S. Deny, M. Chalk, G. Tkačik, O. Marre, T. Mora: *Separating intrinsic interactions from extrinsic correlations in a network of sensory neurons*, Physical Review E, 98 (2018) 042410.

[11] S. Tanaka, H. A. Scheraga: *Model of protein folding: incorporation of a one-dimensional short-range (ising) model into a three-dimensional model*, Proceedings of the National Academy of Sciences, 74 (1977) 1320.

[12] A. Imparato, A. Pelizzola, M. Zamparo: *Ising-like model for protein mechanical unfolding*, Physical Review Letters, 98 (2007) 148102.

[13] Stefan Bornholdt: *Expectation bubbles in a spin model of markets: Intermittency from frustration across scales*, International Journal of Modern Physics C, 12 (2001) 667.

[14] J.-Ph. Bouchaud: *Crises and collective socio-economic phenomena: Simple models and challenges*, Journal of Statistical Physics, 151 (2013) 606|.

[15] D. Sornette: *Physics and financial economics (1776 - 2014): puzzles, ising and agent-based models*, Reports on progress in physics, 77 (2014) 062001.

[16] C. Campajola, F. Lillo, D. Tantari: *Unveiling the relation between herding and liquidity with trader lead-lag networks*, Quantitative Finance, in press, (2020) .

[17] C. Campajola, D. Di Gangi, F. Lillo, D. Tantari: *Modelling time-varying interactions in complex systems: the Score Driven Kinetic Ising Model*, arXiv preprint arXiv:2007.15545 (2020)

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

[19] A. C. Coolen, R. Kühn, P. Sollich: *Theory of neural information processing systems.* Oxford University Press, Oxford (2005).

[20] J. J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities.* Proceedings of the national academy of sciences, 79(8), 2554-2558 (1982).

[21] R. McEliece, E. Posner, E. Rodemich, S. Venkatesh: *The capacity of the Hopfield associative memory*, IEEE transactions on Information Theory, 33 (1987) 461.

[22] G. E. Hinton, T. J. Sejnowski, D. H. Ackley *Boltzmann machines: Constraint satisfaction networks that learn*. Pittsburgh, PA: Carnegie-Mellon University, Department of Computer Science (1984).

[23] D. H. Ackley, G. E. Hinton, T. J. Sejnowski: *A learning algorithm for Boltzmann machines.,* Cognitive science, 9 (1985) 147.

[24] A. Barra, A. Bernacchia, E. Santucci, P. Contucci: *On the equivalence of Hopfield networks and boltzmann machines*, Neural Networks, 34 (2012) 1.

[25] A. Barra, G. Genovese, P. Sollich, D. Tantari: *Phase transitions in Restricted Boltzmann Machines with generic priors*, Physical Review E, 96 (2017) 042156.

[26] E. Agliari, D. Migliozzi, D. Tantari: *Non-convex multi-species Hopfield models* Journal of Statistical Physics 172 (2018) 1247.

[27] G. Genovese, D. Tantari: *Legendre equivalences of spherical Boltzmann machines.* Journal of Physics A: Mathematical and Theoretical, 53 (2020) 094001.

[28] A. Barra, G. Genovese, P. Sollich, D. Tantari: *Phase diagram of restricted Boltzmann machines and generalized Hopfield networks with arbitrary priors.* Physical Review E, 97 (2018) 022310.

[29] A. Decelle, S. Hwang, J. Rocchi, D. Tantari: *Inverse problems for structured datasets using parallel TAP equations and RBM*. arXiv preprint arXiv:1906.11988 (2019).

[30] J. Tubiana, R. Monasson: *Emergence of compositional representations in restricted Boltzmann machines.* Physical Review Letters, 118 (2017) 138301.

[31] D. Krotov, J. J. Hopfield: *Dense associative memory for pattern recognition.* In Advances in neural information processing systems (2016) 1172.

[32] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio: *Deep learning (Vol. 1).* MIT press, Cabridge (MA) (2016).

Daniele Tantari: è Professore Associato di Fisica Matematica presso l'Università di Bologna. Si occupa di Meccanica Statistica e sue applicazioni in Biologia, Intelligenza Artificiale e Finanza.

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

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

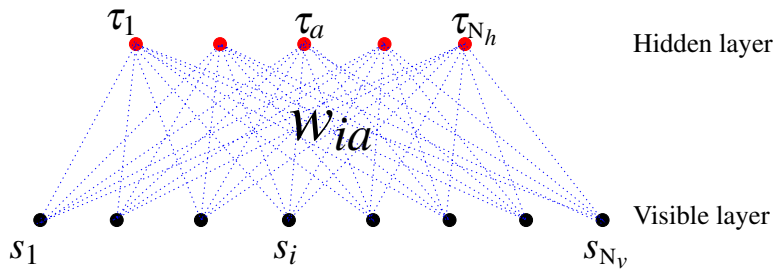


Figura 1: *Struttura bipartita della RBM.*
Bipartite structure of the RBM.

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

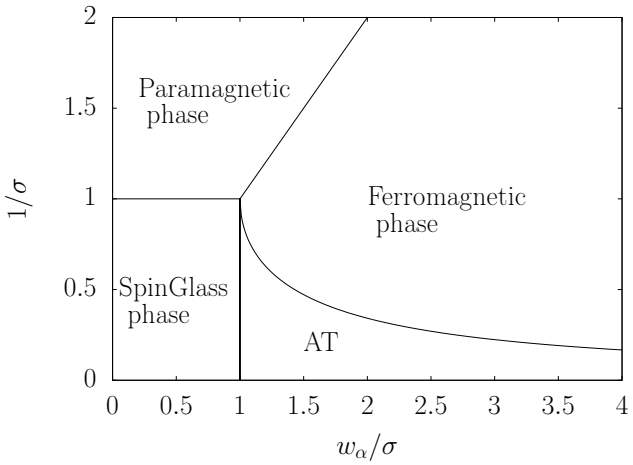


Figura 2: *Diagramma di fase della RBM in funzione del rumore σ e del modo singolare maggiore w_α. Aggiungiamo anche la linea-AT sotto la quale la soluzione simmetrica replica diventa instabile.*
Phase diagram of the RBM as a function of the noise σ and the highest singular mode w_α. We add the AT-line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

Machine Learning: accuratezza, interpretabilità e incertezza

Guido Sanguinetti *School of Informatics, University of Edinburgh, UK; SISSA, Trieste, Italy*

Gli algoritmi di *Machine Learning* sono diventati il motore essenziale dei sistemi di intelligenza artificiale (AI). Nonostante il loro indiscutibile successo, la loro diffusa applicazione a diversi problemi pone molte domande difficili, sia legate al loro impatto sulla società, sia al loro funzionamento tecnico. In questo breve contributo, mi concentro sulla duplice questione della quantificazione dell'incertezza e dell'interpretabilità del modello. Introduco un semplice esperimento teorico per dimostrare che l'importanza di quantificare e scomporre l'incertezza è altamente dipendente dallo scopo del modello, quindi descrivo brevemente i possibili scenari per iniziare ad affrontare questi difficili problemi nell'IA.

Machine Learning algorithms have become the essential engine of artificial intelligence (AI) systems. Despite their unquestionable success, their widespread application to diverse problems poses many difficult questions, both related to their impact on society, and to their technical functioning. In this brief contribution, I focus on the twin issues of uncertainty quantification and model interpretability. I introduce a simple thought experiment to demonstrate that the importance of quantifying and decomposing uncertainty is highly task dependent, and I then briefly describe possible frameworks to start addressing these difficult issues in AI.

Machine Learning: una disciplina in subbuglio

L'ultimo decennio ha assistito ad un'accelerazione esponenziale del progresso nell'intelligenza artificiale

Machine Learning: a discipline in turmoil

The last decade has witnessed an exponential acceleration of progress in Artificial Intelligence (AI).

(IA). Molte pietre miliari apparentemente irraggiungibili sono state conquistate e superate: gli algoritmi di intelligenza artificiale surclassano gli esseri umani in giochi come GO e ATARI [1, 2], svolgono un ruolo centrale in molti processi economici dal settore manufatturiero al settore bancario e stanno facendo grandi passi in avanti nelle scienze, dalla fisica alla biologia. Forse l'*exploit* più impressionante è stata la pubblicazione nel settembre 2020, da parte del quotidiano britannico The Guardian, di un articolo d'opinione interamente scritto da un algoritmo IA [3] L'argomento dell'articolo era una domanda che è sempre più nella mente di molti, scienziati e laici allo stesso modo: "Dobbiamo temere l'ascesa dell'IA?"

Al centro di questo progresso spettacolare c'è l'apprendimento automatico (i.e. *machine learning*, ML). Il ML non è una disciplina nuova: una delle sue principali conferenze, la International Conference on Machine Learning, ha raggiunto la sua trentasettesima edizione nel 2020. Per la maggior parte della sua storia piuttosto oscura, il ML si è concentrato sullo sviluppo di algoritmi informatici che possono essere ottimizzati in ragione dei dati a loro suppliti. Un semplice esempio potrebbe essere l'apprendimento di un classificatore per le immagini associate alle cifre: si cerca una mappa che colleghi i valori di *input* (intensità dei *pixel* in scala di grigi, nel caso delle immagini), con il valore di *output* (l'etichetta della cifra). Il classificatore ha generalmente molti parametri che possono essere regolati in modo tale da minimizzare una funzione di errore su un cosiddetto *training set*, ossia un sottoinsieme dei dati che viene mostrato all'algoritmo; le prestazioni dell'algoritmo vengono quindi valutate empiricamente su un altro set di punti sperimentali, il *test set*, una procedura che può essere giustificata come un'approssimazione Monte Carlo delle prestazioni sui dati estratti dalla effettiva (ma ignota) distribuzione generatrice degli esempi.

Gran parte della ricerca degli ultimi tre decenni è stata dedicata a due compiti principali: sviluppare modi efficaci per riassumere i dati attraverso l'estrazione e la selezione delle loro caratteristiche e lo sviluppo di algoritmi efficienti per eseguire l'inferenza statistica o l'ottimizzazione delle stesse caratteristiche derivate. Esempi del primo tipo di attività includono la vasta letteratura sulla visione artificiale sul rilevamento dei bordi e l'estrazione di caratteristiche, o il campo altrettanto sviluppato dell'analisi grammaticale nell'elaborazione del linguaggio naturale. Esempi del secondo tipo di attività includono algoritmi avanza-

Many seemingly unattainable milestones have been achieved and surpassed: AI algorithms have comprehensively defeated humans at games such as GO and ATARI [1, 2], play a central role in many economic processes from manufacturing to banking, and are making major inroads in sciences from physics to biology. Perhaps most remarkably, in September 2020 the leading UK newspaper The Guardian published an opinion piece entirely authored by an AI algorithm [3] The topic of the article was a question that is increasingly in the mind of many, scientists and lay people alike: "Should we fear the rise of AI?"

At the core of this spectacular progress is Machine Learning (ML). ML is not a new discipline: one of its premiere conferences, the International Conference on Machine Learning, reached its thirty-seventh edition in 2020. For most of its rather obscure history, ML has been focussed on developing computer algorithms which can be tuned by observing data. A simple example could be learning a classifier for digits: one seeks a map that connects input values (pixel intensities in grayscale, in the case of digits), with the output value (the digit label). The classifier generally has many parameters which can be tuned in such a way as to minimise an error function on a so-called *training set*, a subset of the data which is shown to the algorithm; the performance of the algorithm is then empirically evaluated on a separate set of data points, the *test set*, a procedure which can be justified as a Monte Carlo approximation of the performance on data drawn from the true (unknown) data generating distribution.

Much of the research of the last three decades has been dedicated to two major tasks: developing effective ways to summarize data through feature extraction and selection, and developing efficient algorithms to perform statistical inference or optimisation on the features derived. Examples of the first type of task include the vast literature in computer vision on edge detection and feature extraction, or the equally developed field of grammatical parsing in natural language processing. Examples of the second type of task include variational and advanced Markov chain Monte Carlo methods for Bayesian infer-

ti di Markov Chain Monte Carlo per l'inferenza bayesiana, o metodi di ottimizzazione e proiezione casuale per le *kernel machines*. Come molti altri campi scientifici, in particolare nell'informatica, la ricerca in ML è stata condotta principalmente in un'industria artigianale composta da piccoli gruppi formati da un ricercatore principale ed alcuni studenti. Il principale interesse industriale per il ML era molto raro, e in effetti Microsoft era probabilmente l'unico attore industriale significativo nel ML prima del 2010. Anche la più grande conferenza sul ML, NIPS (ora NeurIPS), raggiungeva a malapena 1000 partecipanti e si svolgeva in un'atmosfera informale simile a un ricongiungimento familiare.

Questo stato di cose è cambiato drasticamente nell'ultimo decennio. Ora, tutte le più grandi aziende tecnologiche impiegano legioni di ricercatori in ML e molti professori hanno lasciato il mondo accademico per lavori più redditizi, potenzialmente con un serio effetto a catena sulla capacità delle università di formare nuovi ricercatori in ML. Forse ancor più impressionante, le aziende utilizzano *software* per registrare i propri dipendenti in massa alle principali conferenze di ML già una frazione di secondo dopo l'apertura della registrazione: in pratica, la registrazione a conferenze come NeurIPS o ICML è diventata impossibile per persone che non stanno presentando un lavoro, chiudendo di fatto il campo agli estranei che, in passato, avrebbero potuto partecipare alla conferenza per imparare un pò di più sul ML. Come è successo tutto questo?

L'ascesa del *deep learning*

Il *deep learning* è universalmente riconosciuto come il fattore scatenante nella rivoluzione del ML [4]. L'idea alla base del *deep learning* è che, proprio come il cervello umano sembra imparare solo da esempi senza essere stato progettato per compiti specifici, anche un algoritmo di apprendimento dovrebbe apprendere in modo puramente basato sui dati, evitando qualsiasi ingegnerizzazione delle caratteristiche dei dati ma semplicemente estraendole dagli stessi. Il deep learning lo fa applicando ripetutamente trasformazioni non lineari adattabili ai dati grezzi (si veda il riquadro reti neurali profonde per una breve introduzione ai concetti salienti a riguardo). I modelli costruiti in questo modo sono spesso altamente parametrizzati (cioè hanno molti più parametri liberi che dati suppliti) e di ottimizzazione molto difficile. Quence, or optimisation and random projection methods for kernel machines. Like many other scientific fields, particularly in computer science, ML research was carried out primarily in a cottage industry of small groups of one PI and a few students. Major industrial interest in ML was very rare, and indeed Microsoft was likely the only significant industrial player in ML before 2010. Even the largest ML conference, NIPS (now NeurIPS), hardly reached 1000 participants and encouraged a family reunion-like informal atmosphere.

This state of affairs changed dramatically in the last decade. Now, all of the largest tech companies employ legions of ML researchers, and many professors have left academia for more lucrative jobs, potentially with a serious knock-on effect on the ability of Universities to train new ML researchers. Perhaps more impressively, companies use software tools to register their own employees en masse onto the premiere ML conferences only a fraction of a second after registration is open. Practically, registering for conferences such as NeurIPS or ICML has become impossible for people who are not presenting a paper, effectively closing the field to outsiders who might once have gone to a conference to learn a bit more about ML. How did this all happen?

The rise of deep learning

Deep learning is universally recognised as the trigger of the ML revolution [4]. The idea behind deep learning is that, just as the human brain seems to learn only from examples without being engineered for specific tasks, so should a learning algorithm also learn in a purely data-driven fashion, avoiding any feature engineering and simply extracting features itself. Deep learning does so by repeatedly applying tuneable non-linear transformations to the raw data (see box Deep Neural Networks for a very brief introduction to the fundamental concepts). The models constructed in this way often are highly over-parametrised (i.e., they have many more tuneable parameters than data) and of very difficult optimisation. This and other problems meant that deep neural networks were rapidly abandoned after a short-lived

sto e altri problemi hanno fatto sì che le reti neurali profonde siano state rapidamente abbandonate dopo una popolarità di breve durata negli anni ottanta.

Verso la fine del primo decennio di questo secolo, l'interesse per le reti neurali profonde è tornato a crescere, principalmente grazie al massiccio aumento della disponibilità di dati per addestrarle e dei significativi progressi nell'*hardware* dei computer (in particolare le unità di elaborazione grafica, GPU) che permettono l'ottimizzazione su larga scala. Nel giro di pochi anni, gli algoritmi di *deep learning* hanno vinto con un margine considerevole molte competizioni di visione artificiale e hanno stabilito nuovi standard nel riconoscimento vocale, nell'elaborazione del linguaggio naturale e nella traduzione automatica, per citare solo alcune delle loro principali aree di applicazione. Ciò ha generato sia un enorme interesse industriale, sia un'esplosione di attività di ricerca che non ha precedenti nella storia dell'informatica o, forse, della scienza.

Interpretabilità e incertezza: sono veramente importanti?

Una delle principali conseguenze del passaggio al *deep learning* è stata la perdita dell'interpretabilità umana degli algoritmi. Mentre una caratteristica dei modelli realizzati a mano era che questi venivano generalmente progettati per riflettere il problema in questione, le reti neurali profonde seguono pedissequamente i dati, rimescolandoli attraverso diverse trasformazioni non lineari e rappresentazioni di apprendimento che –sebbene statisticamente efficaci– di solito non corrispondono a caratteristiche interpretabili dall'uomo. Inoltre, per costruzione, le reti neurali profonde hanno uno spazio molto complesso di configurazioni di parametri quasi equivalenti: questo implica che, anche se fosse possibile identificare una spiegazione plausibile per una previsione, molte altre spiegazioni (corrispondenti a configurazioni di parametri equivalenti) potrebbero essere ugualmente valide.

Strettamente correlato a questo è il problema dell'incertezza: molti approcci statistici classici di ML hanno consentito esplicitamente l'incorporazione dell'incertezza negli algoritmi, propagando il rumore attraverso l'algoritmo e quantificando l'incertezza nel risultato finale. Le reti profonde, d'altra parte, sono progettate per approssimare una funzione de-

popularity in the eighties.

Towards the end of the first decade of the century, interest in deep neural networks was resurgent, primarily due to massive increases in data availability, and significant advances in computer hardware (graphic processing units, GPUs) which could support optimisation on such large-scale tasks. Within a few years, deep learning algorithms were winning by a considerable margin many computer vision competitions, and setting new standards in speech recognition, natural language processing and machine translation, to name only some of the major areas of application. This has engendered both huge industrial interest, and in an explosion of research activity which is unprecedented in the history of computer science or, possibly, science.

Interpretability and uncertainty: do they matter?

A major consequence of the shift towards deep learning has been the loss of human interpretability of the algorithms. While hand-crafted features and models were generally designed to reflect the problem at hand, deep neural networks follow the data, scrambling through several non-linear transformations and learning representations that, while statistically effective, do not usually correspond to human-interpretable features. Additionally, by construction deep neural networks have a very complex landscape of nearly equivalent parameter configurations: this implies that, even if it were possible to identify one plausible explanation for a prediction, many other explanations (corresponding to equivalent parameter configurations) might be equally valid.

Closely related is the issue of uncertainty: many classical statistical ML approaches explicitly enabled the incorporation of uncertainty in the algorithms, propagating noise through the algorithm and quantifying the uncertainty in the final result. Deep networks, on the other hand, are designed to approximate an unknown *deterministic* function of the input, and in

terministica sconosciuta dell'*input*, ed in generale si limitano a predizioni puntuali. Questo è vero sia per l'incertezza sull'*output* finale, sia per l'incertezza sulla configurazione dei parametri che hanno portato a tale previsione: nessuna delle due può essere generalmente quantificata o scomposta lungo fattori contributivi interpretabili.

Tutto questo importa? La questione è aperta al dibattito. Ad esempio, l'UE impone legalmente un livello (minimo) di interpretabilità per qualsiasi utilizzo dell'IA in un'ampia gamma di settori che coinvolgono il benessere umano e la società. Tuttavia, molti sostenitori di alto profilo dell'apprendimento profondo hanno fortemente discusso contro l'interpretabilità, sulla base del fatto che sono preferibili prestazioni migliori. A volte vengono utilizzati esempi medici: preferiamo un medico che può curarci meglio o un medico che può spiegare meglio la motivazione alla base della diagnosi?

Direi che l'importanza dell'interpretabilità dipende in gran parte dal compito dato alla rete. Per espandere il mio argomento, consideriamo le seguenti previsioni fittizie:

1. Il computer dice: "questa immagine contiene un gatto seduto su un tavolo";
2. Il computer dice: "Good Morning = Buongiorno = Guten Morgen";
3. Il computer dice: "in base alla tua genetica ed al tuo stile di vita, la tua aspettativa di vita residua è di 20 anni, 5 mesi e tre giorni".

Questo esempio è chiaramente costruito per essere estremo, con lo scenario 3 radicalmente diverso da 1 e 2; tuttavia, credo che esso evidenzi una serie di questioni su cui penso valga la pena riflettere. Prima di tutto, lo scenario 3 implica una previsione reale su un fatto futuro, al contrario di una previsione statistica, la cui bontà sia eventualmente valutabile su un *test set.* In quanto tale, lo scenario 3 non può essere verificato indipendentemente in alcun modo (se non aspettando i 20 anni prescritti). Quindi, sembra almeno ragionevole aspettarsi di essere in grado di capire come sia stata raggiunta tale conclusione. In secondo luogo, proprio perché lo scenario tre appartiene al futuro, potrebbe essere possibile intervenire per modificare il risultato. Tuttavia, come si può capire quali azioni potrebbero essere più efficaci, se non si sa cosa abbia contribuito al raggiungimento della conclusione? È evidente che sia l'interpretabilità

general limit themselves to point predictions. This is true both for uncertainty on the final predicted output, and uncertainty on the configuration of parameters that led to the prediction: neither generally can be quantified or decomposed along interpretable contributing factors.

Does all of this matter? The question is open for debate. For example, the EU mandates legally a (minimum) level of interpretability for any usage of AI in a wide range of areas involving human welfare and society. However, many high-profile proponents of deep learning have powerfully argued against interpretability, on the grounds that better performance is to be preferred. Medical examples are sometimes used: do we prefer a doctor that can cure you better, or a doctor who can explain better the motivation behind the diagnosis?

I would argue that the importance of interpretability is largely task-dependent. To expand on my argument, let's consider the following fictitious predictions:

1. Computer says: "this image contains a cat sitting on a table";
2. Computer says: "Good morning = Buongiorno = Guten Morgen";
3. Computer says: "Based on your genetics and lifestyle, your remaining life expectation is 20 years, 5 months and three days".

This example is clearly constructed to be extreme, with scenario 3 being radically different from 1 and 2; however, I believe it highlights a number of issues which I think are worth reflecting upon. First of all, scenario 3 involves a real prediction about a fact in the future, as opposed to a statistical prediction to be evaluated as held-out data. As such, scenario 3 cannot be verified independently in any way (except by waiting the prescribed 20 years). Hence, it seems at least reasonable to expect to be able to understand how the conclusion was reached. Secondly, precisely because scenario three is in the future, it may be possible to take action to modify the outcome. However, how can I understand which actions might be most effective, if I don't know what contributed to the conclusion? It is evident that both interpretability *AND* uncertainty quantification are needed whenever rational decision making is involved. Finally, while scenarios 1 and 2

sia la quantificazione dell'incertezza siano entrambe necessarie ogni volta che è coinvolto un processo decisionale razionale. Infine, mentre gli scenari 1 e 2 non hanno un impatto immediato sulla vita di una persona, lo scenario 3 sì, e quindi si può sostenere che la sua importanza diretta per qualcuno sia molto più alta.

Vorrei proporre che queste tre semplici domande possano essere poste ad un qualsiasi tipo di previsione:

1. è probabile che la previsione abbia un impatto significativo sulla vita delle persone?
2. la previsione può essere facilmente convalidata in modo indipendente?
3. può essere intrapresa un'azione per modificare il risultato previsto (se in futuro)?

Si noti che prima domanda è già al centro del requisito legale per la spiegabilità nell'UE. Le risposte a queste tre domande insieme, a mio parere, determinano in gran parte se l'interpretabilità e la quantificazione dell'incertezza sono una bella aggiunta o una componente indispensabile di qualsiasi sistema di IA.

Interpretabilità: due approcci possibili

Naturalmente, non sono il primo a discutere questioni di interpretabilità in ML. In effetti, l'IA spiegabile (XAI) è un'area di ricerca in rapida espansione a sé stante e diverse direzioni di ricerca interessanti sono in fase di sviluppo attivo (vedere ad esempio [9] per una recente indagine sul panorama della ricerca in questo campo). Sebbene sia impossibile rendere giustizia al campo, descriverò brevemente due approcci alternativi al problema.

Metodi *post-hoc*

Diversi metodi mirano a spiegare il processo decisionale mediante algoritmi di apprendimento profondo in modo post-hoc. In pratica, ciò consiste nell'individuare quali caratteristiche iniziali abbiano maggiormente contribuito ad una decisione. Ad esempio, il metodo di *layer-wise relevance propagation* [10] spiega la classificazione di un'immagine monitorando l'impatto che l'attivazione di un singolo pixel ha sulla classificazione finale. In questo modo, il metodo

do not immediately have an impact on a person's life, scenario 3 does, and so one may argue that its direct importance to someone is much higher.

I would like to propose that these three simple questions could be asked of any type of prediction:

1. is the prediction likely to have a significant impact on people's life?
2. can the prediction be easily independently validated?
3. can action be taken to modify the predicted outcome (if in the future)?

Notice that Q1 is already at the core of the legal requirement for explainability in the EU. The answers to these three questions together, in my opinion, largely determine whether interpretability and uncertainty quantification are a nice addition or an indispensable component of any AI system.

Interpretability: two possible approaches

Naturally, I am not the first to discuss issues of interpretability in ML. In fact, explainable AI (XAI) is a rapidly expanding research area in its own right and several interesting research directions are being actively developed (see e.g. [9] for a recent survey of the research landscape in this field). While it is impossible to do justice to the field, I will briefly describe two alternative approaches to the problem.

Post-hoc methods

Several methods aim at explaining decision making by deep learning algorithms in a post-hoc way. In practice, this consists in identifying which initial features contributed most to a decision. For example, the method of *layer-wise relevance propagation* [10] explains the classification of an image by tracking the impact that an individual pixel activation has on the final classification. In this way, the method can highlight (for example by colouring them) the most

può evidenziare (ad esempio colorandoli) i *pixel* più rilevanti di un'immagine, dando una spiegazione intuitiva dell'origine del risultato finale. Naturalmente, l'idea non è specifica per le immagini, ma applicabile a qualsiasi *input* ad alta dimensionalità.

Il principale punto di forza di questa classe di metodi è che non compromettono le prestazioni, poiché vengono applicati dopo che l'algoritmo di ML è stato addestrato. Tuttavia, rimangono molte sfide aperte: innanzitutto sembra dubbioso che tali metodi consentano agli utenti di comprendere l'incertezza nelle previsioni, poiché vengono semplicemente applicati post-hoc su un algoritmo pre-addestrato (deterministico). In secondo luogo, le caratteristiche rilevanti sono raramente facilmente riconducibili a categorie comprensibili dall'uomo. Ad esempio, due immagini di cani potrebbero essere entrambe classificate correttamente a causa di sottoinsiemi di *pixel* completamente diversi, –uno a causa della coda e l'altro a causa delle orecchie– rendendo difficile trovare una spiegazione generale di cosa sia un cane secondo l'algoritmo di ML. Ancora più importante, modificare leggermente l'immagine di un cane potrebbe cambiare completamente l'insieme di caratteristiche che l'algoritmo ritiene essere rilevanti per il suo riconoscimento. Ciò ha un impatto drammatico sulla terza domanda della sezione precedente: come possiamo agire riguardo a una previsione, quando la spiegazione offerta è fragile alle perturbazioni?

Interpretabile per costruzione: modelli bayesiani gerarchici

Un'alternativa naturale è progettare algoritmi che siano interpretabili direttamente, ricollegandosi essenzialmente a ciò che i modellisti statistici hanno fatto per decenni. Una struttura particolarmente attraente è fornita dai modelli bayesiani gerarchici (HBM, vedere ad esempio [11]). Tali modelli scompongono esplicitamente l'incertezza lungo componenti ampie e gerarchicamente organizzate. Questa idea è al centro di molte applicazioni mediche stratificate: ad esempio, la variazione totale di un determinato biomarcatore all'interno di una popolazione potrebbe essere scomposta in diverse fonti parzialmente annidate, a partire dalla variazione dovuta a sesso, etnia, fattori di stile di vita e terminando con intrinseca variabilità a livello del singolo individuo. Questi modelli, formulati in termini di probabilità condizionali e solitamente addestrati tramite inferenza sulla

relevant pixels in an image, giving an intuitive explanation for the origins of the final result. Naturally, the idea is not specific to images, but could be applied to any high-dimensional input.

The major strength of this class of methods is that they do not compromise performance, as they are applied after the ML method has been trained. Nevertheless, several open challenges remain. First of all, it seems dubious that such methods would enable users to understand uncertainty in predictions, since they are simply applied post-hoc to a pre-trained (deterministic) method. Secondly, the relevant features are rarely easily relatable to human-understandable categories. For example, two images of dogs could be both correctly classified due to completely different subsets of pixels, one because of the tail, and the other because of the ears, making it difficult to find a general explanation of what is a dog according to the ML algorithm. More importantly, slightly modifying the image of one dog could completely change the set of features that the algorithm finds to be relevant. This impacts dramatically on the third question in the previous section: how can we act about a prediction, when the explanation offered is fragile to perturbations?

Interpretable by design: hierarchical Bayesian models

A natural alternative is to design algorithms to be interpretable directly, essentially reconnecting with what statistical modellers have been doing for decades. A particularly attractive framework is provided by hierarchical Bayesian models (HBMs, see for example [11]). Such models explicitly decompose uncertainty along broad, hierarchically organised components. This idea is at the core of many stratified medicine applications: for example, the total variation in a certain biomarker within a population might be decomposed across several partially nested sources, starting from variation due to gender, ethnicity, lifestyle factors, and ending with intrinsic variability at the level of the single individual. These models, formulated in terms of conditional probabilities and usually trained via posterior inference, precisely quantify and decompose uncertainty in predictions, and are still very

distribuzione a posteriori, quantificano e scompongono con precisione l'incertezza nelle previsioni e sono ancora ampiamente utilizzati in particolare nelle applicazioni mediche.

Naturalmente, tali modelli possono essere utilizzati solo laddove esiste una solida conoscenza di base del fenomeno in esame, poiché sono intrinsecamente sviluppati su misura ed adattati alle applicazioni specifiche. Non avrebbe molto senso sviluppare HBM per immagini naturali, dove non c'è una chiara comprensione del processo generativo sottostante. Ancora più importante, gli HBM sono spesso vincolati a forme funzionali parametriche e relativamente semplici, a causa delle difficoltà computazionali nell'esecuzione dell'inferenza bayesiana in modelli complessi. Tuttavia, recenti sviluppi –ad esempio nelle applicazioni alla genomica *high-throughput* [12, 13]– stanno iniziando ad allentare questi vincoli, introducendo dipendenze funzionali non lineari, guidate dai dati all'interno dei modelli e sfruttando strumenti avanzati di ML come inferenza variazionale stocastica [14].

Discussione

Le tecnologie di Intelligenza Artificiale alimentate da algoritmi di *machine learning* hanno già avuto un impatto enorme sulla nostra vita e sulla nostra società e probabilmente continueranno a farlo nel prossimo futuro. Dati gli enormi livelli di investimento nelle tecnologie da parte di attori sia pubblici che privati è chiaro che il loro utilizzo diventerà sempre più diffuso. È quindi essenziale che potenziali aree problematiche in questa tecnologia, o nella sua applicazione, siano identificate e corrette tempestivamente.

In questo breve contributo, ho tentato di spiegare come la mancanza di spiegabilità e quantificazione dell'incertezza siano, a mio parere, aspetti estremamente problematici delle attuali tecnologie che derivano intrinsecamente dalla progettazione interamente data-driven dei moderni algoritmi di ML. Tali problemi sono fondamentalmente alla radice anche di molti altri problemi, ampiamente pubblicizzati, connessi all'uso delle tecnologie di ML. L'ingiustizia, ad esempio, è una conseguenza quasi inevitabile della mancanza di interpretabilità: è estremamente difficile implementare le nozioni di equità in algoritmi che si basano interamente su dati presi da una società ingiusta. Un altro problema importante è la difficoltà nell'adattare gli algoritmi a cambiamenti (noti) delle condizioni al contorno: mentre per gli algoritmi fatti

widely employed particularly in medical applications.

Naturally, such models can only be used where there is a solid knowledge base, as they are intrinsically bespoke and tailored to specific applications. It would not make much sense to develop HBMs for natural images, where there is no clear understanding of the underlying generative process. More importantly, HBMs are often constrained to parametric and relatively simple functional forms, due to computational difficulties in performing Bayesian inference in complex models. However, recent developments, for example in applications to high-throughput genomics [12, 13], are starting to relax these constraints, introducing non-linear, data-driven functional dependencies within models and taking advantage of advanced ML tools such as stochastic variational inference [14].

Discussion

AI technologies powered by ML algorithms have already massively impacted our life and society, and are likely to continue to do so in the foreseeable future. Given the enormous levels of investment by both public and private actors in the technologies, it is clear that their usage will become more and more widespread. It is therefore essential that potential fault lines either in the technology or in its application are identified and corrected early.

In this brief contribution, I have attempted to explain how lack of explainability and uncertainty quantification are, in my opinion, hugely problematic aspects of current technologies which arise intrinsically from the entirely data-driven design of modern ML algorithms. Such issues are fundamentally at the root also of many other, widely publicised problems with ML technologies. Unfairness, for example, is an almost unavoidable consequence of lack of interpretability: it is extremely difficult to implement notions of fairness in algorithms which are entirely reliant on data taken from an unfair society. Another major issue is the difficulty in adapting algorithms to (known) changes in conditions: while for handcrafted algorithms this is in principle straightforward, as they usually contain explicit, knowledge derived models of how external covariates affect predictions,

a mano questo è in linea di principio semplice, poiché di solito tali algoritmi contengono modelli espliciti e derivati, (anche) da tale conoscenza, per molte `scatole nere` di impiego nel ML, la soluzione è raccogliere un nuovo *training set*, cosa che potrebbe essere irrealizzabile.

La comunità ML è ben consapevole di questi problemi ed in effetti sono stati fatti diversi tentativi per risolverli. Nella parte finale di questa relazione ho evidenziato due linee di ricerca diverse ma complementari a tal proposito. Entrambe sono oggettivamente lontane dall'aver quadrato l'intrattabile cerchio del mantenimento dell'accuratezza fornendo interpretabilità e quantificazione dell'incertezza, ma sono, a mio parere, passi preziosi nella giusta direzione.

Tuttavia, è importante sottolineare che la maggior parte dei metodi di ML, inclusi tutti quelli qui descritti, sono di natura fondamentalmente correlativa: i modelli sono individuati nei dati associati ai risultati di interesse, ma la ricostruzione del flusso causale rimane oltre il loro scopo. Incorporare idee e metodi del ML negli approcci di inferenza causale [15] è certamente una grande sfida per il futuro dell'IA

for many black-box ML approaches the solution is to gather a new training set, which may be infeasible.

The ML community is well-aware of these problems, and indeed several attempts are being made to address them. In the final part of this report, I have highlighted two different but complementary lines of research. Both are objectively far from having squared the intractable circle of retaining accuracy while providing interpretability and quantifying uncertainty, but they are, in my opinion, valuable steps in the right direction.

Nevertheless, it is important to underline that most ML methods, including all the ones described here, are fundamentally correlative in nature: patterns are spotted in the data which are associated with outcomes of interest, but reconstructing the causal flow remains beyond their scope. Incorporating ideas and methods from ML in causal inference approaches [15] is certainly a grand challenge for the future of AI.

Reti neurali profonde

Come funzionano le reti neurali profonde? Le reti profonde sono costituite da un gran numero di cosiddetti `neuroni`, organizzati in `strati`, collegati tra loro in modo gerarchico e *feed-forward*. I dati di *input* sono essi stessi visti come una collezione di attivazioni neuronali: tipicamente, ad ogni dimensione dell'input corrisponde un neurone e per convenzione lo strato di input è considerato il primo livello. Nella sua forma più elementare, ogni neurone nello strato t riceve segnali di *input* da (un sottoinsieme) di neuroni nello strato $t-1$, e quindi esegue le seguenti operazioni: combina linearmente questi segnali (utilizzando un insieme di coefficienti regolabili chiamati `pesi sinaptici`), quindi applica una funzione non lineare $\sigma\colon \mathbb{R} \to \mathbb{R}$ allo scalare risultante.

Deep neural networks

How do deep neural networks work? Deep networks are made up of large numbers of so-called `neurons` which are organised in `layers`, connected to each other in a hierarchical, feed-forward manner. Input data points are themselves seen as a collection of neuronal activations: typically, each input dimension corresponds to one neuron, and by convention the input layer is considered the first layer. In its most basic form, each neuron in layer t receives inputs from (a subset) of neurons at layer $t-1$, and then performs the following operations: it combines linearly the inputs (using a set of tuneable coefficients called `weights`), and then applies a nonlinear function $\sigma\colon \mathbb{R} \to \mathbb{R}$ to the resulting scalar.

Possono essere usati diversi tipi di non linearità (le cosiddette `unità`): i primi tentativi usavano unità sigmoidali come la tangente iperbolica, mentre ad oggi si preferiscono le unità lineari rettificate (i.e. ReLU). È importante sottolineare che le unità devono essere differenziabili quasi ovunque affinchè si possano applicare metodi di ottimizzazione basati sul calcolo del gradiente. In sintesi, ogni neurone i allo strato t produce un *output*

$$o_{it} = \sigma\left(\sum_{j \in \mathscr{J}_{it}} w_{it} o_{j(t-1)}\right)$$

dove $\mathscr{J}_{it}$ è l'insieme di neuroni allo strato $t-1$ che alimentano il neurone i allo strato t. Lo strato finale implementa il classificatore: nel caso di un classificatore binario, tale strato è costituito da un singolo neurone che esegue una regressione lineare generalizzata e fornisce un numero compreso tra 0 e 1. Da questa prospettiva, si possono immaginare gli strati intermedi come dispositivi efficaci che permettono di far imparare alla rete nel suo insieme una `rappresentazione` dei dati ottimale, che li renda separabili da un classificatore lineare. L'apprendimento viene eseguito tramite la discesa del gradiente su una funzione di errore scelta in modo appropriato (tipicamente si usa la *cross-entropy* per la classificazione). Indicata come $y_i \in \{0,1\}$ l'etichetta associata all'istanza di addestramento i e come $f(\mathbf{x}_i, \mathbf{w}) \in [0,1]$ la previsione ad essa associata della rete neurale, la funzione da minimizzare è

$$L(Y,X,W) = \sum_{i \in \mathscr{I}} \Big[y_i \log\left(f(\mathbf{x}_i, \mathbf{w})\right) + (1-y_i)\left(\log\left(1 - f(\mathbf{x}_i, \mathbf{w})\right)\right) \Big]$$

dove $\mathscr{I}$ è l'insieme di indici delle istanze di addestramento e X, Y e W indicano –collettivamente e rispettivamente– dati di input per l'addestramento, relative etichette di addestramento ed i pesi della rete. Poiché le unità sono scelte per essere differenziabili quasi ovunque, ne consegue che anche la funzione da minimizzare è quasi ovunque differenziabile, quindi la minimizzazione può essere ottimizzato rispetto ai pesi mediante metodi di discesa del gradiente.

Different types of nonlinearity can be used (so called `units`): early efforts used sigmoidal units such as hyperbolic tangent, while more recently rectified linear units are preferred. Importantly, units need to be differentiable almost everywhere to apply gradient-based optimisation methods. In summary, each neuron i at layer t produces an output

$$o_{it} = \sigma\left(\sum_{j \in \mathscr{J}_{it}} w_{it} o_{j(t-1)}\right)$$

where $\mathscr{J}_{it}$ is the set of neurons at layer $t-1$ feeding into neuron i at layer t. The final layer implements the classifier: in the case of a binary classifier, it consists of a single neuron performing generalised linear regression and outputting a number between 0 and 1. In this sense, one can view the purpose of the intermediate layers as a device to learn a `representation` of the data that makes it optimally separable by a linear classifier. The learning is performed by gradient descent on an appropriately chosen error function (typically, the *cross-entropy* loss function for classification). Denoting as $y_i \in \{0,1\}$ the label associated with training instance i, and as $f(\mathbf{x}_i, \mathbf{w}) \in [0,1]$ the associated prediction of the neural network, the loss function is

$$L(Y,X,W) = \sum_{i \in \mathscr{I}} \Big[y_i \log\left(f(\mathbf{x}_i, \mathbf{w})\right) + (1-y_i)\left(\log\left(1 - f(\mathbf{x}_i, \mathbf{w})\right)\right) \Big]$$

where $\mathscr{I}$ is the index set of the training instances and X, Y, and W denote collectively training inputs, training labels and weights of the network. Since the units are chosen to be differentiable almost everywhere, it follows that the loss function is also a.e. differentiable, and can be optimised w.r.t. the weights by gradient descent methods.

I gradienti vengono calcolati automaticamente usando la differenziazione simbolica e, per ottenere la scalabilità, vengono usati sottoinsiemi casuali di dati (detti *minibatches*) usati per calcolare prontamente ogni passo del gradiente, portando a una `discesa stocastica lungo il gradiente`, che appare anche più efficace nell'evitare minimi locali. In pratica, i minibatch non vengono scelti del tutto casuale ma mediante un protocollo che assicuri che l'intero set di addestramento venga utilizzato iterativamente attraverso quella che viene chiamata `epoca` dell'ottimizzazione.
Le reti neurali profonde possono essere pensate come uno schema di approssimazione di funzioni basato su funzioni di base adattive; è noto da molto tempo che le reti profonde possono approssimare arbitrariamente bene qualsiasi funzione regolare [5]. Ma come possono funzionare in pratica? Dopo tutto, per costruzione, le reti profonde hanno un numero molto elevato di pesi e un numero quasi infinito di simmetrie (ad esempio, la permutazione dei neuroni all'interno di uno strato dovrebbe portare a soluzioni identiche). Empiricamente, la presenza di molti ottimi locali o l'overfitting non sembrano essere veri problemi (sebbene le moderne reti neurali siano piene di trucchi euristici per evitarli). Recenti lavori teorici hanno anche dimostrato che, in un limite termodinamico adeguato, il problema degli ottimi locali è naturalmente evitato e la convergenza globale è ottenuta dalla discesa stocastica lungo il gradiente [6, 7, 8].

Gradients are computed automatically using symbolic differentiation, and, in order to achieve scalability, random subsets of the data (*minibatches*) are used to compute each gradient step, leading to a `stochastic gradient descent`, which also appears more effective in avoiding local optima. In practice, minibatches are not chosen entirely at random but in a schedule that ensures that the whole training set is used iteratively through what is called an `epoch` of the optimisation.
Deep neural networks can be thought of as a function approximation scheme based on adaptive basis functions; it has been known for a long time that deep networks can approximate arbitrarily well any smooth function [5]. But how can it possibly work in practice? After all, by construction deep networks have a very large number of weights and a nearly infinite number of symmetries (e.g., permuting neurons within a layer should lead to identical solutions). Empirically, local optima or overfitting do not appear to be problems (although modern neural networks are replete with heuristic tricks to avoid such problems). Recent theoretical work has also shown that, in a suitable thermodynamic limit, local optima issues are naturally avoided, and global convergence is achieved by stochastic gradient descent [6, 7, 8].

[1] V. Mnih, et al.: *Human-level control through deep reinforcement learning*, Nature 518 (2015) 529.

[2] D. Silver, et al.: *Mastering the game of go without human knowledge*, Nature 550 (2017) 354.

[3] https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

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

[5] G. Cybenko: *Approximation by superpositions of a sigmoidal function*, Math. of Contr. Sign. and Sys. 4 (1989), 303.

[6] G.M. Rotskoff, E. Vanden-Eijnden: *Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error*, arXiv preprint arXiv:1805.00915 (2018).

[7] S. Mei, A. Montanari, P.M. Nguyen: *A mean field view of the landscape of two-layer neural networks*, Proc. Natl. Acad. Sci. USA, 115 (2018) E7665.

[8] S.S. Du et al., *Gradient descent finds global minima of deep neural networks*, arXiv preprint arXiv:1811.03804 (2018).

[9] W. Samek, et al., *Explainable AI: interpreting, explaining and visualizing deep learning*, W. Samek, G. Montavon, A. Vedaldi, L. K Hansen, K.-R. Müller (Eds.) Springer-Nature, Berlin (2019) 16.

[10] S. Bach et al.: *On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation*, PLoS One 10 (2015) e0130140.

[11] A. Gelman et al., *Bayesian data analysis*, CRC Press, New York (2013).

[12] N. Eling et al., *Correcting the mean-variance dependency for differential variability testing using single-cell RNA sequencing data*, Cell systems, 7 (2018) 284.

[13] C.A. Kapourani, R. Argelaguet, G. Sanguinetti, C.A. Vallejos, *scMET: Bayesian modelling of DNA methylation heterogeneity at single-cell resolution*, bioRxiv, Cold Spring Harbor Laboratory (2020).

[14] R. Ranganath, S. Gerrish, D. Blei, *Black box variational inference*, Proc. of Seventeenth International Conference on Artificial Intelligence and Statistics, PMLR (2014) 814.

[15] J. Pearl, *Bayesianism and causality, or, why I am only a half-Bayesian*, Foundations of bayesianism (2001) 19.

Guido Sanguinetti: è professore di Fisica Applicata ed ha la Cattedra di Data-Science all'International School for Advanced Studies (SISSA) in Trieste. È inoltre Professor of Computational Bioinformatics alla School of Informatics, University of Edimburgh, dove ha insegnato dal 2010. I suoi interessi risiedono nel machine learning applicato a sistemi biomedicali, in particolare problemi di systems biology ed high-throughput biology.

Guido Sanguinetti: is Professor of Applied Physics and Chair of Data Science at the International School for Advanced Studies (SISSA) in Trieste. He is also Professor of Computational Bioinformatics at the School of Informatics, University of Edinburgh, UK, where he has worked since 2010. His interests are in machine learning applied to biomedical systems, in particular problems in systems and high-throughput biology.

Piccole reti neurali crescono

Carlo Lucibello *Bocconi Institute for Data Science and Analytics, Università Bocconi, Milano, Italy*

Diamo qui un breve un accenno della fenomenologia delle *deep neural networks*. Vediamo poi come delle piccole reti, trattabili analiticamente attraverso le tecniche della fisica statistica, siano in grado di catturare un parte di questa fenomenologia.

Il *Loss landscape*

Le reti neurali artificiali sono indubbiamente la forza motrice che sta dietro al rapido sviluppo a cui abbiamo assistito negli ultimi 10 anni nel campo dell'intelligenza artificiale.

Numerosi sono i traguardi recentemente conseguiti che, oltre a motivare il sempre crescente interesse accademico, hanno affascinato, stupito e a volte forse anche intimorito il grande pubblico.

`Alpha-Go`, un algoritmo basato su *deep reinforcement learning* e su `Monte-Carlo` *tree search*, è riuscito nell'impresa che si riteneva ancora molto lontana di battere il campione mondiale di `Go`, Lee Sedol. Sedol, in seguito alla sconfitta, ha abbandonato il gioco competitivo. `Alpha-Go` ha poi lasciato spazio ad `Alpha-Zero`, una rete in grado di allenarsi da zero, ovvero senza supervisione umana ma solo competendo con se stessa.

`GPT-3`, un modello di generazione linguistica con 175 miliardi di parametri, è in grado di generare testi difficilmente distinguibili da quelli creati da uno scrittore umano.

Modelli generativi come le *Generative Adversarial Networks* (`GANs`) sono in grado di produrre immagini e video estremamente realistici. Questi modelli sono comunemente impiegati nella creazione di quelli che vengono chiamati i `Deep Fake`.

Nonostante le architetture neurali si facciano via via più complesse e diversi moduli vengano combinati in maniera sorprendente grazie alla fantasia dei ricercatori, la matematica ed i principi di funzionamento alla base rimangono molto semplici.

In via piuttosto generale, una rete neurale può essere pensata come una relazione tra *input* e *output*, ovvero una funzione, formata da una serie di operazioni algebriche seguite da operazioni *element-wise* non-lineari. Più precisamente, un rete neurale è caratterizzata da un insieme di parametri che chiameremo θ nel seguito, per cui è in grado di esprimere un'intera famiglia di funzioni $\{f_\theta\}$ al variare di tali parametri. L'**apprendimento** consiste proprio nel selezionare all'interno di questo vasto spazio dei parametri una configurazione che ben si adatti ai dati osservati.

Una caratteristica che ha contribuito al successo delle reti neurali, ed in particolare di quelle *deep*, e che le rende uno strumento molto versatile, è la loro espressività, ovvero la capacità di ben approssimare una arbitraria relazione di *input-output*.

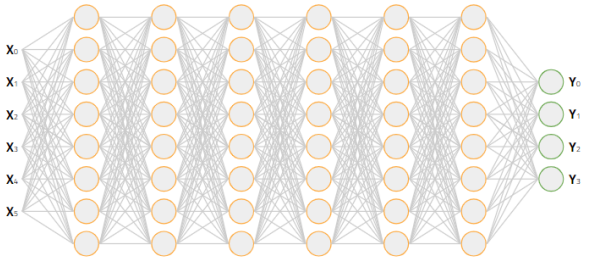


Figura 1: *Rappresentazione grafica di un multi-layer pereptron*

Il multilayer perceptron. Un esempio classico di *deep neural network* è il `multi-layer perceptron (MLP)`, rappresentato in Fig. 1. Un MLP con L *layers* è costituito da una matrice W_ℓ (i pesi sinaptici) e da un vettore $\boldsymbol{b}_\ell$ (il bias) per ogni *layer* ℓ. Inoltre ad ogni *layer* è associata una funzione σ_ℓ che opera *element-wise* e che viene detta funzione di attivazione. Per ogni dato vettore di *input* $\boldsymbol{x}$, la rete fornirà in *output* una predizione $\hat{y}$ ottenuta attraverso il cosiddetto **forward pass**:

$$\begin{aligned} \boldsymbol{x}_1 &= \sigma_1(W_1\,\boldsymbol{x} + \boldsymbol{b}_1) \\ \boldsymbol{x}_2 &= \sigma_2(W_2\,\boldsymbol{x}_1 + \boldsymbol{b}_2) \\ &\vdots \\ \hat{y} &= \sigma_L(W_L\,\boldsymbol{x}_{L-1} + \boldsymbol{b}_L) \end{aligned}$$

Chiamando θ l'insieme dei pesi e dei *bias*, abbiamo così definito la nostra funzione di predizione $\hat{y} = f_\theta(\boldsymbol{x})$. In uno scenario di apprendimento supervisionato, si ha a disposizione un *dataset* D di esempi costituiti da coppie di *input* $\boldsymbol{x}$ ed etichette y. Il problema dell'apprendimento è formulato come problema di ottimizzazione di una **Loss function** $\mathcal{L}(\theta)$ che assume la forma

$$\mathcal{L}(\theta) = \mathbb{E}_{(\boldsymbol{x},y)\sim D}\ \ell\left(y, f_\theta(\boldsymbol{x})\right) \tag{1}$$

Qui la funzione $\ell(y,\hat{y})$ è tipicamente l'errore quadratico medio nei *task* di regressione e la *cross-entropy* nei *task* di classificazione.

Ottimizzare la Loss. L'ottimizzazione viene conseguita attraverso varianti del semplice algoritmo iterativo noto come **gradient descent** (GD):

$$\theta^{t+1} = \theta^t - \eta\nabla\mathcal{L}(\theta^t) \tag{2}$$

Difatti, nonostante una pletora di metodi più sofisticati, quali ad esempio la famiglia dei metodi di `quasi-Newton`, siano stati prodotti dalla comunità di ottimizzazione, l'altissima dimensionalità del problema rende computazionalmente proibitivo ogni approccio significativamente più complesso del semplice GD. Inoltre, per guadagnare in efficienza computazionale e far fronte all'enorme aumento della dimensione dei *dataset* usati per il *training*, una variante dello schema di GD, nota come **Stochastic Gradient Descent** (SGD) viene di fatto utilizzata. Nello schema di SGD, un sottoinsieme casuale degli esempi B_t, detto **mini-batch**, viene estratto ad ogni iterazione ed usato per calcolare una versione approssimata della `Loss` (1) che a sua volta fornirà una approssimazione del gradiente usato nell'update di Eq. (2). La dinamica di SGD è quindi definita da:

$$\theta^{t+1} = \theta^t - \eta\nabla\mathcal{L}_{B_t}(\theta^t) \tag{3}$$

Alla fine del `training`, la rete viene testata su un *set* di esempi escluso dal *training set*, in modo da valutare l'effettiva capacità della rete di generalizzare, ovvero rispondere correttamente in situazioni nuove ma simili a quelle di cui si è fatta esperienza. Questo è il criterio che anche in termini comuni useremmo per giudicare un apprendimento di successo. Ed è proprio qui, nella straordinaria capacità di generalizzare, nonostante il grande numero di parametri che caratterizzano questi modelli, che sta un'altra delle chiavi del successo del `deep learning`.

Ora che abbiamo definito il nostro *setting*, è interessante fare una serie di considerazioni.

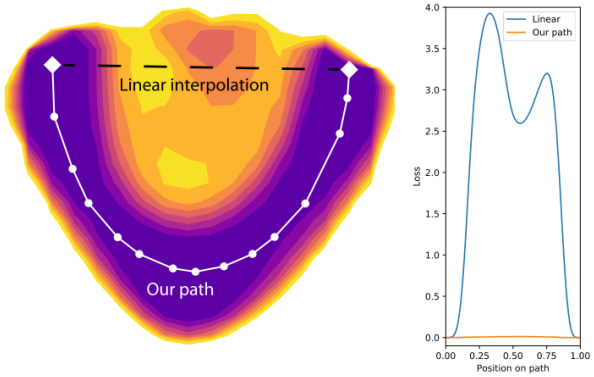
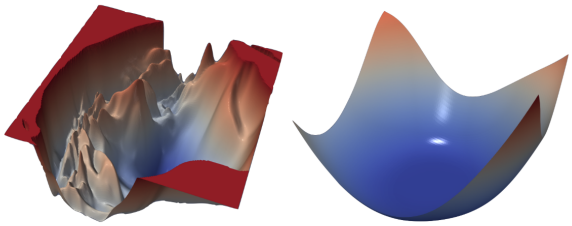


Figura 2: *Proiezione in due dimensioni di un percorso a bassa training loss che connette due soluzioni algortimiche. Immagine da Ref. [1]*

Minimi locali, minimi globali. La `Loss` di Eq. (1) è generalmente una funzione altamente non-convessa. Il suo *landscape* potrebbe quindi essere molto complesso e possedere un numero molto grande di minimi locali e di selle, anche a valori di `loss` alti. L'evidenza empirica però dimostra che reti sufficientemente grandi e allenate con SGD arrivano facilmente a valori di loss molto bassi. Non rimangono quindi incastrati in minimi locali. Mentre questo fenomeno era stato inizialmente attribuito allo stocasticità intrinseca di SGD, potenzialmente in grado di permettere al sistema di evadere dai minimi locali, l'evidenza corrente è che questa sia una proprietà propria del *landscape* e delle condizioni di inizializzazione. Se minimi locali alti esistono, non sembrano essere algoritmicamente rilevanti. Inoltre, per alcune architetture e sotto assunzioni molto forti e spesso irrealistiche, si hanno dei risultati che escludono la presenza di minimi locali che non siano anche globali.

Un fondo connesso e rugoso. L'evidenza sperimentale suggerisce che i minimi algoritmicamente accessibili possono essere connessi lungo traiettorie a bassa `loss` (si veda Fig. 2, sinistra). Generalmente questi percorsi sono tortuosi: una semplice interpolazione lineare tra due minimi darà l'impressione di trovarsi di fronte a delle barriere (Fig. 2, destra).



(a) ResNet-110, no skip connections (b) DenseNet, 121 layers

Figura 3: *L'uso di connessioni residue contribuisce a smussare e appiattire il landscape della training loss. Immagine da Ref. [2].*

Minimi larghi, minimi stretti, generalizzazione. Scelte architetturali, quali ad esempio il numero di layer, la loro larghezza, l'impiego di connessioni residue, hanno un enorme impatto sulla geometria della `loss`, come ad esempio mostrato in Fig. 3. Alcuni iperparametri che regolano il *training*, in particolare la *batch size*, il *learning rate* e l'utilizzo del *dropout*, influenzano la piattezza della `loss` intorno alle soluzioni trovate. Ciò ha un impatto rilevante sulle proprietà di generalizzazione. L'evidenza numerica mostra che minimi larghi hanno migliori proprietà di generalizzazione di quelli stretti.

La dinamica. L'analisi della dinamica di GD, e ancor di più di SGD, è sostanzialmente intrattabile, complicata dall'alta dimensionalità e dalla generale non-convessità del problema. SGD è in prima approssimazione simile ad una dinamica di Langevin, seppur vi entri un rumore non-bianco e dipendente dalla posizione. Misure sperimentali delle funzioni di correlazione a due punti nella dinamica del *training* ci mostrano delle connessioni con le dinamiche lente proprie dei sistemi fisici vetrosi [3].

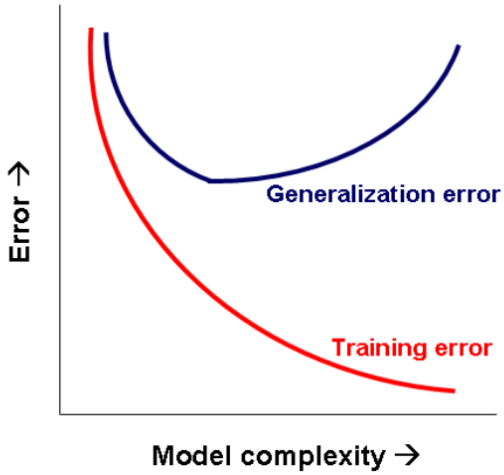


Figura 4: *Il problema dell'overfitting nella teoria statistica classica*

La teoria classica dell'apprendimento statistico. La teoria classica dell'apprendimento statistico è in forte difficoltà nello spiegare il successo delle reti neurali. Infatti, tale teoria prevede che quando il numero di parametri N del modello diventa molto maggiore del numero dei dati di training P, $N \gg P$, si va incontro al fenomeno dell'*overfitting* (si veda Fig. 4). Si dovrebbe quindi osservare un aumento delle errore di generalizzazione all'aumentare di N. In pratica invece questo non si asserva per le `deep networks`. Queste tipicamente operano nel regime $N \gg P$,

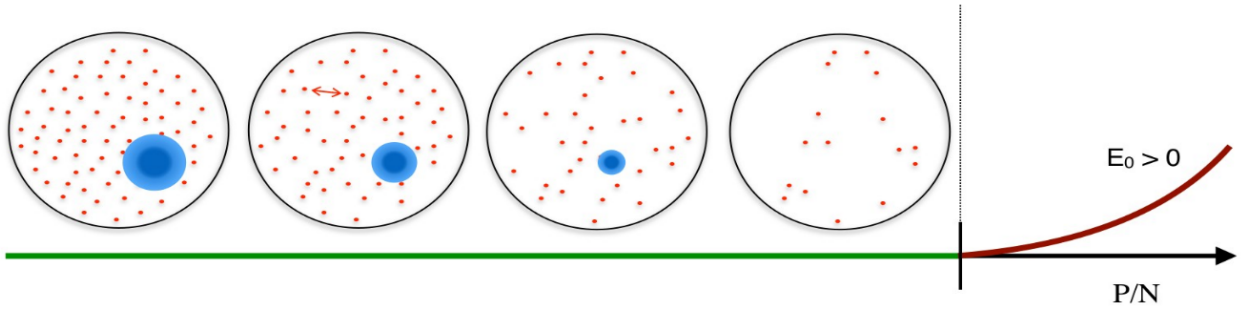


Figura 5: *Lo spazio delle soluzioni nel percettrone binario. La gran parte delle soluzioni sono isolate, ma un numero minore (ma pur sempre esponenziale) forma una regione densa e connessa rappresentata in blu.*

dove i *bound* classici sull'errore di generalizzazione diventano vacui. Tra le sfide concettuali più rilevanti in questo campo vi è il capire la natura di questo fenomeno, la sua relazione con la scelta delle architetture, l'inizializzazione delle reti e la dinamica del *training*. Di enorme rilevanza è l'individuare dei *bound* per l'errore di generalizzazione al tempo stesso più stringenti e computazionalmente trattabili . Il sacro Graal qui consiste nel rendere esplicita una qualche regolarizzazione implicita che agisce in questi modelli e che ne riduce il numero di parametri efficaci, comprimendo lo spazio delle ipotesi.

Il percettrone binario

Dopo questo *excursus* nel mondo delle reti neurali profonde, ci spostiamo ora su uno scenario molto più ristretto, quello della rete neurale minimale, ovvero il singolo neurone, chiamato in questo contesto **percettrone**. Il *task* è quello di classificazione binaria, in due classi che denotiamo con $y = +1$ e $y = -1$. La funzione di predizione associata al percettrone è quindi

$$\hat{y} = \text{sign} \sum_{i=1}^{N} W_i \, x_i^\mu \tag{4}$$

Se i pesi W_i sono lasciati liberi di assumere valori arbitrari nel continuo, si può dimostrare che il problema, in presenza di un *training set* linearmente separabile, ammette un insieme di soluzioni convesso, quindi poco interessante e non rappresentativo di quanto avviene in reti più grandi. Consideriamo invece il caso in cui i pesi possono assumere solo valori binari, $W_i \in \{-1, +1\}$. Andremo a vedere che questo sistema è dotato di una fenomenologia molto ricca. Computazionalmente, il problema diventa più duro (seppur ancora algoritmicamente risolvibile [6]), in quanto non ci si potrà avvalere degli strumenti di ottimizzazione basati sul gradiente. Questo modello, chiamato **percettrone binario**, è stato analizzato per lungo tempo dalla comunità di fisica statistica [4, 5].

Più in generale, la possibilità di allenare reti neurali binarie è argomento di vivace ricerca nella comunità del *machine learning*, in quanto ha notevoli ricadute in termini di efficienza computazionale ed energetica, nonché ai fini della compressione e dell'utilizzo di rete neurali su piccoli dispositivi con scarsa risorse di memoria e computazionali.

Proseguiamo la nostra analisi considerando un *setting* molto semplice, in cui P esempi di *training* sono forniti alle reti. Un esempio è formato da una coppia x^μ e y^μ generata in modo casuale, con un *output* scorrelato dall'*input*. Non c'è quindi in questo caso una regola da imparare (ne forniremo invece una nella prossima sezione) ma siamo solamente interessati ad analizzare il numero di soluzioni di questo *constraint satisfaction problem*. Tale numero è dato dalla funzione di partizione Z associata al problema, definita da

$$Z = \sum_{\boldsymbol{W}} \prod_{\mu=1}^{P} \mathbb{I}\left(y^\mu \sum_{i=1}^{N} W_i \, x_i^\mu > 0 \right) \tag{5}$$

Usando il metodo delle repliche ed il metodo della cavità della teoria dei vetri di spin, è possibile calcolare l'entropia media del sistema nel limite

termodinamico:

$$S = \lim_{N\to\infty} \frac{1}{N} \mathbb{E} \log Z \tag{6}$$

Qui l'aspettazione è sulle realizzazione del *training set* ed il limite viene preso andando a guardare al regime non triviale in cui il numero di esempi rimane proporzionale ad N, ovvero fissiamo una costante α per cui $P = \alpha N$. Il calcolo dell'entropia S, unita ad un'analisi delle *performance* degli algoritmi, rivela una serie di nette transizioni di fase al variare di α in corrispondenza di alcuni valori che chiamiamo α_{alg} e α_c.

- Per $\alpha < \alpha_{alg}$ esiste quasi sempre un numero esponenziale di soluzioni ed è algoritmicamente semplice trovare una di queste. Come vedremo in seguito però, non tutte le soluzioni sono algoritmicamente accessibili (infatti gran parte non lo sono). Il valore di α_{alg} dipende dall'algoritmo considerato.

- Per α intermedio invece, $\alpha_{alg} < \alpha < \alpha_c$, nonostante la presenza di un numero esponenziale di soluzioni, nessun algoritmo noto riesce a raggiungere una soluzione in tempo polinomiale. In questo regime il problema è `duro`.

- Per $\alpha > \alpha_c$ il problema con alta probabilità non ammette soluzioni. Qualsiasi configurazione W classifica male un sottoinsieme degli esempi del *training set*.

Si veda Fig. 5 per una rappresentazione del *set* di soluzioni al variare di α.

Lo scenario Insegnante-Allievo

Arricchiamo ora il problema considerando un modello di generazione dei dati in cui gli *input* $\boldsymbol{x}^\mu$ continuano ad essere *random*, ma in cui le etichette y^μ sono generate da un secondo percettrone, preso con pesi *random*, che chiameremo `l'Insegnante`. Il percettrone su cui viene fatto l'apprendimento sarà chiamato lo `Studente`.

Anche questo contesto è analizzabile attraverso le tecniche della teoria degli *spin-glass*. Possiamo a questo punto porci due domande, strettamente connesse tra di loro: quanto è bravo lo `Studente` a generalizzare, ovvero a classificare correttamente nuovi esempi prodotti dall'`Insegnante`? Siamo capaci di inferire, visti $P = \alpha N$ esempi, il vettore di pesi dell'`Insegnante`?

La risposta dipende dal regime in α in cui operiamo. Ancora una volta abbiamo una sequenza di transizioni, a dei valori di α che chiamiamo α_{IT} e α_*.

- Per $\alpha < \alpha_{IT}$ esiste un numero esponenziale di soluzioni al problema, tra cui l'`Insegnante` stesso. In questo mare di soluzioni non è possibile discriminare l'`Insegnante`. In questo regime l'inferenza perfetta è impossibile dal punto di vista della teoria dell'informazione. Si nota in questo regime una discrepanza tra l'errore di generalizzazione predetto dalla teoria per una soluzione tipica (campionata uniformemente a caso) e quello fornito dalle soluzioni trovate dagli algoritmi, essendo quest'ultimo sistematicamente migliore. Scopriremo nella prossima sezione il perché.

- Per α tale che $\alpha_{IT} < \alpha < \alpha_*$, l'Insegnante si trova ad essere l'unica soluzione, quindi in linea di principio sarebbe possibile farne l'inferenza perfetta e arrivare ad errore di generalizzazione nullo. Tuttavia non sono noti algoritmi efficienti in grado di trovare l'Insegnate, il problema è algoritmicamente duro.

- Per $\alpha > \alpha_*$ l'`Insegnante` si trova ad essere l'unica soluzione e l'inferenza perfetta è ottenibile in tempo polinomiale. Il problema è semplice in questo regime.

Recentemente questo scenario è stato rigorosamente comprovato, confermando le predizioni delle teoria delle repliche [7].

La geometria dello spazio delle soluzioni

In entrambi i problemi analizzati, la descrizione *coarse grained* data dallo studio dell'entropia del sistema, seppur in grado di mettere alla luce delle sorprendenti transizioni di fase, non è in grado di spiegare appieno le *performance* degli algoritmi e la natura delle soluzioni trovate. Quello che

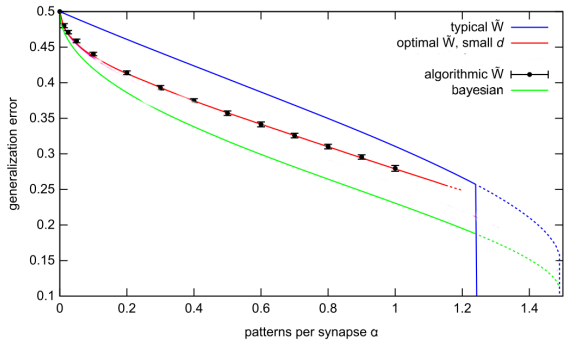


Figura 6: *Errore di generalizzazione teorico predetto per le soluzioni isolate del percettrone binario, comparato a quello delle soluzioni all'interno del cluster denso e a quello dato da soluzioni trovate da algoritmi polinomiali. Questi ultimi due coincidono perfettamente.*

succede infatti, è che le soluzioni trovate dagli algoritmi polinomiali quali `Belief Propagation` (BP), ovvero le soluzioni algoritmicamente accessibili, non corrispondono a soluzioni estratte uniformemente a caso dall'insieme delle soluzioni, ma piuttosto da una misura con un forte *bias* verso alcune regioni.

Un'idea più precisa possiamo farcela andando a vedere nel dettaglio la geometria locale dello spazio delle soluzioni, grazie ad una tecnica proposta da Parisi e Franz nel contesto degli *spin-glass* [8].

Chiamiamo entropia locale $S_{loc}(\tilde{\boldsymbol{W}}, d)$ di una configurazione $\tilde{\boldsymbol{W}}$, il (log-)numero di soluzione del problema entro una distanza intensiva d da $\tilde{\boldsymbol{W}}$, ovvero

$$S_{loc}(\tilde{\boldsymbol{W}}, d) = \frac{1}{N} \log \left[\sum_{\boldsymbol{W}} \prod_{\mu=1}^{P} \mathbb{I}\left(y^\mu \sum_{i=1}^{N} W_i \, x_i^\mu > 0 \right) \times \mathbb{I}\left(\|\boldsymbol{W} - \tilde{\boldsymbol{W}}\| < dN \right) \right] \tag{7}$$

Il calcolo dell'entropia locale per il percettrone, portato avanti in Ref. [9], svela una sorprendente geometria e ci permette di chiarificare le *performance* algoritmiche. Risulta infatti che la gran parte delle soluzioni del sistema è isolato, ovvero si trova ad essere a distanza estensiva da altre soluzioni. Affianco a queste soluzioni dominanti, un numero sempre esponenziale ma sottodominante è raggruppato in un cluster ad alta entropia locale, denso e connesso. Evidenze analitiche e numeriche mostrano che questo *cluster* smette di esistere in corrispondenza delle transizioni algoritmiche (si veda Fig. 5). In altre parole, nei regimi sotto α_c e sotto α_{IT} nei due scenari esaminati, gli algoritmi trovano soluzioni in tempo polinomiale fintanto che il *cluster* denso esiste, e le soluzioni che trovano vivono appunto all'interno di una regione densa di soluzioni. Le soluzioni isolate risultano algoritmicamente inaccessibili.

Abbiamo quindi un *landscape* simile a quello di un campo da golf, arricchito però dalla presenza di una grossa e profonda buca che rende possibile l'apprendimento.

Altro risultato dell'analisi è che le soluzioni all'interno del *cluster* denso generalizzano meglio delle soluzioni isolate (Fig. 6). Abbiamo quindi ritrovato in questo semplice modello l'analogo del *gap* di generalizzazione che si osserva tra minimi larghi e stretti nelle `deep network`.

Conclusioni

In questo articolo, abbiamo brevemente riassunto alcuni tratti fenomenologici che si osservano empiricamente nelle `deep network`, con particolare attenzione al *landscape* della `Loss function` e alle proprietà di generalizzazione dei diversi minimi algoritmicamente accessibili.

Abbiamo poi analizzato un modello di rete neurale analiticamente trattabile, il percettrone binario, ed evidenziato la stretta connessione tra *performance* algoritmiche e geometria dello spazio delle soluzioni. L'intuizione che ne deriva è estendibile al contesto del `deep learning`, ed ha portato all'elaborazione di algoritmi che per costruzione sono indirizzati verso minimi larghi e ad alta entropia locale, con buone proprietà di generalizzazione [10, 11].

Sebbene questo tipo di analisi su modelli semplici porti alle creazione di framework interpretativi molto generali ed anche a ricadute pratiche nel contesto deep, alcuni limiti sono evidenti. Le complesse architetture odierne sono largamente fuori scala per le tecniche di analisi teorica della fisica statistica. Inoltre le assunzioni di dati i.i.d. comunemente usate sono fortemente irrealistiche. Su quest'ultimo fronte, degli importanti passi avanti si sono visti di recente con l'introdu-

zione di modelli generativi di dati strutturati ma analiticamente trattabili [12].

Le piccole reti neurali sono cresciute a dismisura nel corso degli anni, c'è bisogno di nuovi strumenti di analisi teorica per continuare a fare strada assieme.

[1] F. Draxler et al.: *Essentially No Barriers in Neural Network Energy Landscape*, Proceedings of the 35th International Conference on Machine Learning, PMLR 80 (2018) 1308.

[2] H. Li et al.: *Visualizing the Loss Landscape of Neural Nets*, Advances in Neural Information Processing Systems 31 (NIPS 2018) (2018).

[3] M. Baity-Jesi et al. *Comparing dynamics: Deep neural networks versus glassy systems*, Proceedings of the 35th International Conference on Machine Learning, PMLR 80 (2018) 314.

[4] W. Krauth, M. Mézard: *Storage capacity of memory networks with binary couplings*, Journal the Physique, 50 (1989) 3057.

[5] A. Engel, C. Vand der Broeck: *Statistical mechanics of learning* Cambridge University Press, Cambridge (2001).

[6] A. Braunstein, R. Zecchina: *Learning by message passing in networks of discrete synapses*, Physical Review Letters, 96 (2016) 030201.

[7] J. Barbier et al.: *Optimal errors and phase transitions in high-dimensional generalized linear models*, PNAS, 116 (2019) 5451.

[8] S. Franz, G. Parisi: *Recipes for metastable states in Spin Glasses*, Journal de Physique, 5 (1995) 1401.

[9] C. Baldassi et al.: *Subdominant Dense Clusters Allow for Simple Learning and High Computational Performance in Neural Networks with Discrete Synapses*, Physical Review Letters, 115 (2015) 128101.

[10] P. Chaudhari et al. : *Entropy-SGD: Biasing Gradient Descent Into Wide Valleys*, ICLR 2017., arXiv:1611.01838.

[11] C. Baldassi et al.: *Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes*, „ PNAS (113) 2016.E7655

[12] S. Mei, A. Montanari: *The generalization error of random features regression: Precise asymptotics and double descent curve*, arXiv:1908.05355

Carlo Lucibello: è Assistant Professor presso l'università Bocconi. Si occupa di problemi all'interfaccia tra il machine learning e la fisica statistica.

La Rilevanza nell'Apprendimento Statistico

Vivere è imparare

Konrad Lorenz

Matteo Marsili *The Abdus Salam International Centre for Theoretical Physics, Trieste, Italy*

Nonostante lo straordinario sviluppo dell'intelligenza artificiale, la nostra comprensione teorica dell'apprendimento statistico, che ne è alla base, è ancora insufficiente. Offrirò qui una prospettiva basata sul concetto di rilevanza. La rilevanza misura la quantità di informazione sul processo generativo contenuto in un campione di dati o nella rappresentazione interna di un sistema che impara. Spiegherò perché le proprietà di sistemi che imparano differiscono marcatamente da quelle tipiche di sistemi fisici. In particolare, la criticalità emerge come una caratteristica generica di dati massimamente informativi o di sistemi efficienti per l'apprendimento. L'esposizione che segue è discorsiva. Si rimanda agli articoli originali per una trattazione più estesa.

La rivoluzione digitale e l'avvento dell'era dei *Big Data*, hanno portato l'interesse della comunità scientifica sul problema dell'apprendimento statistico automatico, o *machine learning*. Quest'ultimo è alla base degli sviluppi più sorprendenti in Intelligenza Artificiale, poiché la capacità di apprendere è una componente fondamentale di ogni comportamento intelligente. A dispetto delle sue spettacolari applicazioni, dalle traduzioni automatiche alla sconfitta del campione di Go da parte di *AlphaGo*, il nostro grado di comprensione di questi algoritmi è probabilmente paragonabile a quanto si sapeva del motore a vapore nei primi decenni della rivoluzione industriale. Gli ingredienti necessari per realizzare motori potenti (pistoni, pulegge, vapore, ecc.) erano noti ben prima che fosse raggiunta una comprensione soddisfacente del fenomeno. Questa comprensione giunse grazie al lavoro di Carnot, su macchine ideali, che rivelò l'esistenza di una funzione di stato — l'entropia — che gioca un ruolo fondamentale nei processi termodinamici. Allo stesso modo, gli ingredienti per l'apprendimento statistico (ad esempio *algorithms, architectures, and structured data* secondo [1]) sono noti, ma il nostro livello di comprensione è tuttora insoddisfacente.

Infatti, anche se sistemi esperti come quelli di traduzione automatica hanno prestazioni compa-

rabili a quelle dell'intelligenza umana, i dati su cui sono allenati sono generati dall'Uomo. Inoltre, queste prestazioni sono limitate a compiti ben specifici e richiedono costi energetici ordini di grandezza maggiori di quelli impiegati dal cervello umano. Infine, l'inferenza statistica negli organismi viventi deve essere efficiente in regimi in cui i dati sono estremamente scarsi. Le falene che riescono a inferire la posizione della femmina più efficientemente, dalle deboli tracce di feromoni che questa rilascia nell'aria, trasmetteranno più probabilmente i propri geni. Dunque, comprendere l'apprendimento statistico dal punto di vista teorico non è solo importante per le applicazioni di intelligenza artificiale, ma anche per comprendere i sistemi viventi.

Gli strumenti matematici a nostra disposizione per affrontare questi problemi sono gli stessi su cui si basa la statistica classica. Le ultime decadi sono state testimoni di una notevole convergenza tra discipline diverse, dalla teoria dell'informazione, dei codici e della complessità, alla probabilità, la statistica e la meccanica statistica. Per affrontare il problema dell'apprendimento in regimi rilevanti per l'intelligenza artificiale, i concetti della statistica classica devono essere declinati in nuovi ambiti, in cui i dati e i modelli hanno una dimensionalità comparabile se non maggiore al numero di campioni disponibili, ed in cui non abbiamo idea di quale sia il vero modello generativo.

Campioni massimamente informativi: Rilevanza e criticalità

Quanta informazione contiene un campione statistico sul processo generativo sottostante? Nella statistica classica, quando il modello generativo appartiene ad una famiglia parametrica di distribuzioni, questa domanda trova una risposta nel concetto di statistiche sufficienti. Queste sono le osservabili che possiamo calcolare dai dati, e che permettono di stimare i parametri del modello. Tutto il resto dell'informazione contenuta nel campione è rumore non informativo. Possiamo quantificare quanto impariamo, calcolando l'informazione mutua tra il campione ed i parametri del modello. E troviamo che *i)* il numero di bit che impariamo cresce solo con il logaritmo del numero N di dati[1], una frazione risibile dell'informazione contenuta nei dati, che corrisponde allo spazio necessario a immagazzinare i dati sul nostro computer (che cresce con N). Inoltre, *ii)* impariamo quel numero di bit sia che il modello stimato sia quello giusto o meno, o che i dati siano informativi o totalmente aleatori. Per avere garanzie che quei bit che acquisiamo svelino davvero informazioni sul processo generativo, dobbiamo comparare modelli diversi in modo Bayesiano. Tuttavia, il numero dei modelli che dovremmo considerare cresce più che esponenzialmente con la dimensionalità dei dati. Ciò rende questo approccio impraticabile (ma si veda [2]).

In assenza di informazioni *a priori* su quale sia il modello generativo, la massima di Socrate, secondo cui l'unica cosa che possiamo sapere è "saper di non sapere", ci viene in soccorso. Infatti, mentre è difficile identificare una struttura precisa nei dati, la mancanza assoluta di informazione ha una sua matematizzazione precisa nel principio di massima entropia. Questo permette di quantificare esattamente la parte non informativa di un campione in quei sotto-campioni di osservazioni che si presentano lo stesso numero di volte. Il resto dell'informazione, che corrisponde all'entropia di Shannon della distribuzione delle frequenze, è un limite superiore della quantità di informazione che il campione contiene sul modello generativo. Questa entropia, che chiameremo **rilevanza** in quel che segue, è una grandezza fondamentale che ci permette di stimare quanto il nostro campione possa essere informativo. Nel linguaggio comune, il termine rilevanza è spesso inteso in senso relativo. Qui viene inteso in senso assoluto, o meglio relativo al processo generativo dei dati, che ci è ignoto. Per spiegare in modo colorito, ciò che è rilevante per dei neuroni di un'area del cervello o per proteine coinvolte nella stessa funzione biologica, è ciò di cui i neuroni o le proteine parlerebbero al bar, davanti ad un caffè.

Questa caratterizzazione matematica ci permette di identificare campioni di dati massimamente informativi in quelli che hanno massima

[1] Il motivo è che l'inferenza ci permette di ridurre l'incertezza su ogni parametro del modello di un fattore $1/\sqrt{N}$, ed il numero di bit necessari per rappresentare un numero a quella precisione cresce come $\frac{1}{2}\log N$.

rilevanza. Questo problema è ben posto, una volta che consideriamo campioni con lo stesso livello di **risoluzione**. La risoluzione è la quantità di bit necessaria ad immagazzinare un dato del campione sul nostro computer, ed è misurata dall'entropia di Shannon dei dati. La risoluzione può essere scelta dall'osservatore[2]. Ma una volta determinato il livello di dettaglio, la rilevanza è una proprietà intrinseca dei dati, che ci svela informazioni sulla natura del processo generativo.

La figura 1 riporta un esempio della dipendenza della rilevanza in funzione della risoluzione, per rappresentazioni del comportamento temporale di $N = 4000$ azioni del New York Stock Exchange (NYSE) estratte con tre diversi algoritmi di *data clustering*. Si rimanda a [3] per ulteriori dettagli. In breve, gli algoritmi classificano le azioni in *clusters* s, la cui grandezza k_s corrisponde al numero di azioni appartenenti al cluster s. La risoluzione e la rilevanza sono date da

$$\hat{H}[s] = -\sum_s \hat{p}_s \log \hat{p}_s, \quad (1)$$

$$\hat{H}[k] = -\sum_k \hat{p}_k \log \hat{p}_k \quad (2)$$

rispettivamente, dove $\hat{p}_s = k_s/N$ e $\hat{p}_k$ sono le probabilità empiriche che un'azione presa a caso appartenga al *cluster* s, e che tale cluster contenga $k_s = k$ azioni, rispettivamente. Al variare del numero di *clusters* considerati, $\hat{H}[s]$ e $\hat{H}[k]$ tracciano una curva sul piano. Tale curva fornisce una misura di quanto la rappresentazione che diversi algoritmi di *data clustering* estraggono dai dati sono informative sul processo generativo. Se esaminiamo i tre diversi algoritmi riportati nella figura (si veda [3]), scopriamo che l'algoritmo che realizza valori più elevati di $H[k]$ è basato su un modello più aderente alla dinamica delle azioni finanziarie rispetto agli altri, e si avvicina più degli altri alla "vera" classificazione, che è quella data dalla *Security and Exchange Commission* del *New York Stock Exchange*.

In Fig. 1 distinguiamo due regimi. Uno in cui $\hat{H}[k]$ cresce con la risoluzione $\hat{H}[s]$. Questo è il regime in cui ogni stato s è campionato un numero sufficiente di volte e possiamo aspettarci che la distribuzione empirica fornisca una approssimazione ragionevole della vera distribuzione $p(s)$. Ciò non vale per la parte destra, in cui $\hat{H}[k]$ decresce con $\hat{H}[s]$. Questo è il regime di cui parlavamo sopra, in cui la statistica classica non è applicabile e in cui operano l'intelligenza artificiale e quella naturale.

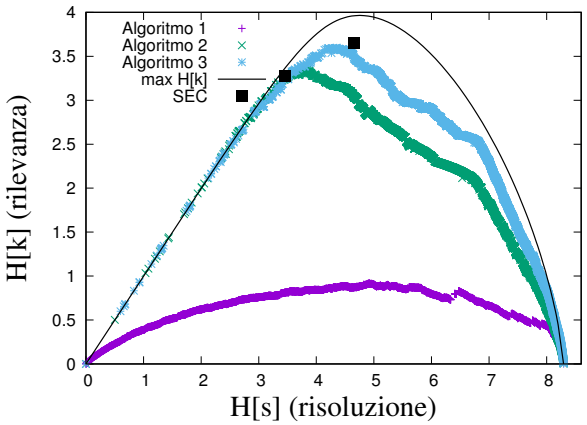


Figura 1: *Andamento di $H[k]$ in funzione di $H[s]$ per tre diversi algoritmi di data clustering, applicati a dati finanziari (si veda [3] per dettagli). Il massimo teorico è riportato dalla curva continua. I due punti ■ corrispondono alla classificazione delle azioni della Security and Exchange Commission (SEC).*

La stessa figura riporta anche la curva corrispondente al valore massimo di $\hat{H}[k]$ per un dato valore di $\hat{H}[s]$. La caratteristica dei campioni che realizzano tale massimo è quella di esibire criticalità statistica, cioè

$$\hat{p}_k \sim k^{-2-\mu}. \quad (3)$$

L'esponente $-\mu$ corrisponde, approssimativamente alla pendenza della curva $\hat{H}[k]$ in funzione di $\hat{H}[s]$. Ciò significa che quando comprimiamo i dati[3], riducendo $\hat{H}[s]$ di un bit, la rilevanza $\hat{H}[k]$ aumenta di μ bit. Possiamo quindi distinguere ulteriormente due sotto-regimi. Quando $\mu > 1$ l'ulteriore compressione rivela una quantità di informazione $\Delta\hat{H}[k]$ maggiore di quella $(-\Delta H[s])$ eliminata. Quando invece $\mu < 1$ la compressione avviene a spese di informazione potenzialmente rilevante. Il punto $\mu = 1$ caratterizza dunque quei campioni massimamente informativi con un livello ottimale di compres-

[2] Ad esempio, possiamo descrivere lo stato di una molecola a diversi livelli di dettaglio, dalla particella puntuale alla descrizione meccanico quantistica in termini di numeri quantici.

[3] Che corrisponde a ridurre il numero dei clusters, nell'esempio di Fig.1.

sione. In questo punto, la statistica di frequenze Eq. (3) corrisponde alla celebre legge di Zipf [4]. Alla luce di quanto detto, l'osservazione della legge di Zipf nella distribuzione delle parole nel linguaggio, o nella frequenza di antigeni nel sistema immunitario – per citare solo due esempi dell'ubiquità di tale legge – può essere letta come conseguenza che sia il linguaggio che il sistema immunitario sono rappresentazioni efficienti per la comunicazione o la difesa dai patogeni, rispettivamente. Più in generale, la prospettiva qui delineata fornisce una chiara risposta alla domanda sul perché la biologia sembri operare in condizioni di criticalità [5]: i sistemi biologici non sono ottimizzati ad un punto critico. Operano su rappresentazioni massimamente informative.

Questa caratterizzazione delle rappresentazioni massimamente informative apre la strada a metodi di inferenza da dati alto-dimensionali *model free*: una direzione di ricerca molto promettente. Ad esempio, [6] mostra come il concetto di rilevanza sia in grado di identificare quei neuroni altamente informativi per la navigazione, unicamente sulla base della loro attività neurale.

Se la criticalià è l'impronta di campioni massimamente informativi, la dovremmo ritrovare nella teoria dei codici, che si occupa appunto di comprimere dati generati da una sorgente in modo efficiente. Questo è stato infatti verificato per codici che realizzano una lunghezza di descrizione minima (*minimum description length*) in [7]. Per questi codici è possibile dimostrare esplicitamente che si tratta di sistemi critici, posizionati esattamente al punto critico di una transizione di fase [7].

La struttura delle macchine ideali per l'apprendimento

Mentre la criticalità è un eccezione per i sistemi fisici, che si verifica solo per valori particolari dei parametri esterni (e.g. temperatura e pressione), la criticalità statistica è la norma per sistemi evoluti al fine di estrarre rappresentazioni efficienti da dati complessi. Per capirlo, dobbiamo aprire la scatola di questi sistemi, e guardare alla distribuzione dei livelli energetici ed alle Hamiltoniane che li caratterizzano. Ci rendiamo allora conto che il problema dell'apprendimento è di natura duale rispetto a quello della meccanica statistica. In entrambi i casi siamo interessati alla statistica degli stati s interni di un sistema in contatto con un sistema $\vec{x}$ esterno. Nella meccanica statistica, s è lo stato di un sottosistema in contatto termico con l'ambiente $\vec{x}$. L'Hamiltoniana $\mathcal{H}$ è data, ed il principio di massima entropia indica nella distribuzione di Gibbs-Boltzmann

$$\arg \max_{p(s):\langle \mathcal{H} \rangle = E} H[s] = \frac{1}{Z} e^{-\mathcal{H}(s)/T}, \qquad (4)$$

il risultato, dove $H[s] = -\sum_s p(s) \log p(s)$ è l'entropia di Shannon. Eq. (4) rende esplicito che l'unica informazione che il sistema fisico ricorda del suo ambiente $\vec{x}$ è l'energia media $\langle \mathcal{H} \rangle$, o equivalentemente la temperatura T, che deve coincidere con quella dell'ambiente[4]. Come conseguenza, la distribuzione dell'energia $\mathcal{H}$ del sistema è concentrata in un piccolo intervallo dell'ordine di $1/\sqrt{n}$ attorno al valor medio $\langle \mathcal{H} \rangle = E$, dove $n \gg 1$ è il numero di gradi di libertà[5]. Questo risultato prende il nome di **Proprietà Asintotica dell'Equipartizione** (PAE) [9] in probabilità. Questa afferma che quasi sicuramente, tutti gli stati osservati hanno la stessa probabilità $p(s) \sim e^{-nH[s]}$ e che, come conseguenza, il numero di tali stati tipici è $\sim e^{nH[s]}$.

In un sistema ottimizzato per apprendere e classificare gli stati $\vec{x}$ dell'ambiente, l'Hamiltoniana assume il naturale significato di costo informativo $\mathcal{H}(s) = -\log p(s)$. Per riuscire a ricordare il massimo numero di dettagli del suo ambiente $\vec{x}$, che in questo caso sono i dati che il sistema vuole imparare, un tale sistema deve lottare contro la PAE, modellando l'Hamiltoniana $\mathcal{H}(s)$ in modo tale che dati $\vec{x}$ e $\vec{x}'$ con una diversa struttura, corrispondano a stati interni s e s' con energie significativamente diverse $\mathcal{H}(s) \neq \mathcal{H}(s')$, come dimostrato in [10]. Ne consegue che i sistemi che estraggono rappresentazioni efficienti dai dati debbano avere una distribuzione di livelli d'energia più larga possibile. Usando ancora l'entropia

[4] Si osservi che $\langle \mathcal{H} \rangle$ è la statistica sufficiente della distribuzione Eq. (4). L'unica variabile necessaria a determinare il parametro coniugato T.

[5] Tale intervallo si allarga solo quando il sistema è al punto critico di una transizione di fase, cioè quando la distinzione tra due diverse fasi termodinamiche scompare. Ad esempio, come mostrato in [8], per il modello di Ising in campo medio, le fluttuazioni dell'energia sono dell'ordine di $n^{-3/4}$ alla temperatura critica.

come misura quantitativa di informazione, ciò porta al **principio di massima rilevanza**

$$\mathcal{H}^*(s) = \arg \max_{\mathcal{H}(s):\ \langle \mathcal{H} \rangle = H[s]} H[E], \qquad (5)$$

dove

$$H[E] = -\sum_E p(E) \log p(E), \qquad (6)$$

e $p(E)$ è la probabilità che uno stato tipico della macchina abbia energia $\mathcal{H}(s) = E$. La rilevanza $H[E]$ misura la quantità di informazione che la rappresentazione s mantiene sul processo generativo dei dati $\vec{x}$, alla risoluzione $H[s]$. A sua volta, $H[s]$ è il costo informativo medio della rappresentazione. Le soluzioni del problema (5) sono caratterizzate da una densità di stati esponenziale $W(E) = \sum_s \delta\left(\mathcal{H}(s) - E\right) \simeq W_0 e^{\mu E}$ o, equivalentemente, da distribuzioni esponenziali dell'energia

$$p(E) = W(E) e^{-E} \simeq e^{(\mu - 1)E}. \qquad (7)$$

Questo è l'analogo della criticalità statistica che caratterizza campioni massimamente informativi[6] (Eq. 3). Il parametro μ gioca ancora lo stesso ruolo. La rilevanza cresce con la risoluzione $H[s]$ quando $\mu \leq 1$. Per $\mu = 1$ la rilevanza $H[E]$ raggiunge il suo massimo, corrispondente ad una distribuzione $p(E)$ approssimativamente uniforme (Eq. 7). Per valori maggiori di μ, $H[E]$ decresce con la risoluzione. Nei termini della differenza

$$H[s|E] = H[s] - H[E] = \langle \log W(E) \rangle, \qquad (8)$$

che fornisce una misura del rumore non-informativo della rappresentazione, è facile comprendere che l'interrelazione tra risoluzione e rilevanza è della stessa natura di quello discusso sopra per un campione: La diminuzione di un bit in $H[s]$ riduce $H[s|E]$ di $\mu - 1$ bit. Per $\mu > 1$ la compressione elimina solo dettagli non informativi della rappresentazione e rivela dettagli rilevanti, mentre per $\mu < 1$ una frazione $1 - \mu$ dell'informazione è persa.

Il confronto tra le equazioni (4) e (5) rende esplicita la relazione duale dei due problemi. Tale dualità si manifesta nel fatto che l'entropia $H[s|E] = \langle \log W(E) \rangle$ è una funzione concava dell'energia $H[s] = \langle E \rangle$ in meccanica statistica, mentre è convessa nell'apprendimento ottimale, cioè per le soluzioni dell'Eq. (5). Inoltre la criticalità si verifica solo in condizioni speciali in meccanica statistica, mentre essa è una caratteristica tipica di macchine allenate nell'apprendimento di strutture complesse di dati. è curioso osservare come "essere critici nell'apprendimento" non sia solo importante nel senso comune del termine, ma anche nell'accezione tecnica che il termine assume in meccanica statistica.

La rilevanza $H[E]$ fornisce anche un limite inferiore alla quantità di informazione che la rappresentazione interna di una macchina che impara estrae sulle caratteristiche nascoste (*hidden features*) dei dati [8]. Le *hidden features* appaiono come le statistiche sufficienti della rappresentazione interna che la macchina genera nell'apprendimento. Massimizzare $H[E]$ implica dunque estrarre il massimo di informazioni possibile sulle *hidden features*.

Il principio di massima rilevanza trova riscontro nel comportamento di macchine reali (*Restricted Boltzmann Machines, Deep Belief Networks, etc.*), come dimostrato in [8, 11]. Dove invece fallisce totalmente è nel caso di macchine di apprendimento Gaussiane, in cui la rilevanza resta costante nel processo di apprendimento, indipendentemente dalla struttura dei dati. Un fallimento illuminante, che rivela la differenza sostanziale tra apprendimento e stima di parametri [8]: non esistono *hidden features* in un modello Gaussiano, dal momento che le statistiche sufficienti sono ben visibili *a priori*. Tale modello impara solo medie e covarianze, ed è cieco rispetto ad ogni struttura dei dati che vada oltre ai primi due momenti.

Questo è un ulteriore indicazione della sostanziale differenza tra la statistica classica, che opera in regimi in cui la complessità dei modelli è limitata dai dati disponibili, e l'intelligenza artificiale dove l'apprendimento si avvale di macchine sovra-parametrizzate, che possono adattare la loro Hamiltoniana per modellare strutture complesse di dati.

[6] Infatti le frequenza k_s/N dello stato s in un campione approssima la probabiliità $e^{-\mathcal{H}(s)}$ di quello stato. Il cambio di variabili $\mathcal{H}(s) \simeq -\log(k_s/N)$ trasforma l'Eq. (7) nell'Eq. (3).

La meccanica statistica di macchine ideali di apprendimento

L'apprendimento genera una distribuzione $p(s, \vec{x})$ i cui massimi descrivono coppie $(s, \vec{x})$ di dati tipici $\vec{x}$ e rappresentazioni interne s corrispondenti. Le proprietà generiche del problema

$$\max_{s,\vec{x}} \log p(s, \vec{x}) = \max_{s,\vec{x}} \left[\mathcal{H}(s) + \mathcal{V}(\vec{x}|s)\right], \qquad (9)$$

possono essere studiate grazie alla teoria delle statistiche degli estremi, a seconda della distribuzione dei livelli di energia del sistema $\mathcal{H}(s)$ e dell'Hamiltoniana di interazione con l'ambiente $\mathcal{V}(\vec{x}|s)$. Tale studio è stato realizzato in Ref. [12], assumendo per entrambi una distribuzione esponenziale *stretched* di parametro γ,

$$P\{\mathcal{H}(s) \geq E\} = e^{-(E/\Delta)^\gamma}, \qquad (10)$$
$$P\{\mathcal{V}(\vec{x}|s) \geq V\} = e^{-V^\gamma}. \qquad (11)$$

Ciò permette di paragonare le proprietà di sistemi fisici, come il celebre *Random Energy Model* (REM) [13] ($\gamma = 2$) a quelle di macchine di apprendimento ideali ($\gamma = 1$). Il parametro Δ in Eq. (10) regola la scala di energia dei livelli della rappresentazione interna $\Delta \sim \mathcal{H}(s)$. Per ogni valore di γ, si osserva una transizione di fase, al variare di Δ, tra uno stato in cui la distribuzione $p(s)$ è concentrata su un numero finito di stati ($\Delta > \Delta_c$), ed uno stato "disordinato" in cui invece $p(s)$ si estende su un numero esponenziale di stati ($\Delta < \Delta_c$). Nell'apprendimento, la fase ordinata ($\Delta > \Delta_c$) corrisponde a rappresentazioni compresse dei dati.

Tale transizione è continua per $\gamma > 1$ (sistemi fisici) e discontinua per $\gamma < 1$. Inoltre, per $\gamma > 1$ la soglia critica Δ_c decresce all'aumentare della dimensionalità relativa del sistema s, rispetto all'ambiente $\vec{x}$, che per sistemi fisici corrisponde al bagno termico. Ciò rende possibile una descrizione statistica di un sistema, indipendentemente dalla conoscenza di tutte le variabili che possono influenzarlo. Ad esempio, se la struttura di una proteina dipendesse da ogni dettaglio delle interazioni con altre molecole presenti nella cellula, non sarebbe possibile inferirne la struttura dalla sequenza di amino-acidi. Tale predizione statistica risulta invece possibile [14]. Per $\gamma < 1$ invece, più fattori influenzano lo stato interno s del sistema, e meno tale stato è predicibile. Inoltre, non è possibile ricostruire l'Hamiltoniana di sistemi con $\gamma < 1$ attraverso un processo di misurazione sperimentale. Ciò significa che i sistemi con $\gamma < 1$ non sono "apprendibili". Ciò conferisce un ruolo speciale alle macchine ideali di apprendimento ($\gamma = 1$) come quelle che separano i sistemi che possiamo conoscere da quelli inaccessibili. Inoltre, solo per $\gamma = 1$ si verifica la situazione speciale in cui Δ_c è indipendente dalla dimensionalità di $\vec{x}$. Per una macchina di apprendimento, in cui $\vec{x}$ corrisponde ai dati ed s alla rappresentazione interna, questa proprietà è fondamentale. Ad esempio, desideriamo macchine che classifichino le immagini nello stesso modo, indipendentemente dalla risoluzione in pixels delle immagini stesse. Solo sistemi con una distribuzione esponenziale di stati ($\gamma = 1$) hanno questa proprietà.

Conclusioni

Le applicazioni di intelligenza artificiale pongono delle domande fondamentali sull'apprendimento e sulla natura delle macchine che estraggono rappresentazioni efficienti dai dati. La fisica statistica ha fornito importanti contributi in questa direzione. Queste pagine suggeriscono che, per sviluppare un quadro concettuale aderente alla realtà di sistemi efficienti di apprendimento, sia necessario comprendere le peculiari differenze tra questi sistemi e i sistemi fisici. Lo sviluppo di macchine di apprendimento più efficienti o di metodi per estrarre variabili rilevanti da dati alto-dimensionali, basate su questo quadro concettuale, sono promettenti direzioni di ricerca.

Più in generale, come osservato da Paul Davies [15], non abbiamo criteri quantitativi per distinguere materia inanimata da forme di vita. Ciò è cruciale, ad esempio, nell'astrobiologia per capire se la vita si è sviluppata anche altrove nell'Universo. Se l'apprendimento è una caratteristica distintiva della vita, il concetto di rilevanza può contribuire concretamente a colmare questa lacuna.

[1] L. Zdeborovà: *Understanding deep learning is also a job for physicists*, Nat. Phys. 16, (2020) 602.

[2] C. de Mulatier, P. P. Mazza, M. Marsili: *Statistical Inference of Minimally Complex Models*, arXiv:2008.00520 [cs.AI] (2020).

[3] M. Marsili, I. Mastromatteo, Y. Roudi: *On sampling and modeling complex systems* Journal of Statistical Mechanics: Theory and Experiment (2013) P09003

[4] G. K. Zipf: *Human Behavior and the Principle of Least Effort* Addison-Wesley, Cambridge, Mass. (1949).

[5] T. Mora, W. Bialek: *Are biological systems poised at criticality?*, Journal of Statistical Physics, 144 (2011) 268.

[6] R. J. Cubero, M. Marsili, Y. Roudi: *Multiscale relevance and informative encoding in neuronal spike trains* Journal of computational neuroscience, 48 (2020) 85.

[7] R. J. Cubero, M. Marsili, Y Roudi: *Minimum description length codes are critical* Entropy 20 (2018) 755

[8] O. Duranthon, M. Marsili, R. Xie: *Maximal Relevance and Optimal Learning Machines* arXiv:1909.12792 (2019)

[9] Thomas M. Cover, Joy A. Thomas: *Elements of information theory* John Wiley & Sons, New York (1999).

[10] R. J. Cubero, J. Jo, M. Marsili, Y Roudi, J. Song: *Statistical criticality arises in most informative representations* Journal of Statistical Mechanics: Theory and Experiment (2019) 063402.

[11] J. Song, M. Marsili, Jungyo Jo: *Resolution and relevance trade-offs in deep learning* Journal of Statistical Mechanics: Theory and Experiment (2018) 123406.

[12] M. Marsili *The peculiar statistical mechanics of optimal learning machines* Journal of Statistical Mechanics: Theory and Experiment (2019) 103401.

[13] B. Derrida, *Random-energy model: An exactly solvable model of disordered systems* Physical Review B 24 (1981) 2613.

[14] F. Morcos, et al. *Direct-coupling analysis of residue coevolution captures native contacts across many protein families* Proceedings of the National Academy of Sciences 108 (2011) E1293

[15] P. Davies, *Does new physics lurk inside living matter?*, Physics Today 73 (2020) 34.

Matteo Marsili: è *Senior Research Scientist* presso l'*Abdus Salam International Centre for Theoretical Physics*. Si occupa di applicazioni interdisciplinari della meccanica statistica.

Inferenza ad alta dimensionalità: una prospettiva di meccanica statistica

Machines take me by surprise with great frequency.
A. Turing

Jean Barbier *The Abdus Salam International Center for Theoretical Physics, Trieste, Italy*

L'inferenza statistica è la scienza del trarre conclusioni inerentemente un sistema utilizzando dati. Nelle moderne tecniche di signal processing e machine learning, l'inferenza viene eseguita in dimensioni molto elevate: moltissime caratteristiche sconosciute del sistema devono essere dedotte da molti dati rumorosi ed in un numero molto alto di dimensioni. Questo "regime ad alta dimensionalità" ricorda la meccanica statistica, che mira a descrivere il comportamento macroscopico di un sistema complesso basandosi sulla conoscenza delle sue interazioni microscopiche. Ad oggi è chiaro che ci sono molte connessioni tra inferenza e fisica statistica. Questo articolo ambisce evidenziare alcuni dei profondi legami che collegano queste discipline, apparentemente separate, at-

Statistical inference is the science of drawing conclusions about some system by using data. In modern signal processing and machine learning, inference is done in very high dimension: very many unknow characteristics about the sytem have to be deduced from a lot of high dimensional noisy data. This "high-dimensional regime" is reminiscent of statistical mechanics, which aims at describing the macroscopic behavior of a complex system based on the knowledge of its microscopic interactions. It is by now clear that there are many connections between inference and statistical physics. This article aims at emphasising some of the deep links connecting these apparently separated disciplines through the description of paradigmatic models of high-dimensional inference in the

traverso la descrizione di modelli paradigmatici di inferenza ad alta dimensionalità nel linguaggio della meccanica statistica.

Inferenza statistica: il vecchio ed il nuovo

L'**inferenza statistica** ambisce descrivere accuratamente un sistema, mediante l'impiego di una distribuzione di probabilità appropriata, basata sui dati relativi a questo sistema e, potenzialmente, su alcune ipotesi ad esso inerenti. La studio di procedure di inferenza che siano tanto statisticamente quanto computazionalmente efficienti è quindi cruciale praticamente in tutti i campi della scienza.

La statistica classica si occupa principalmente del regime in cui il sistema in studio è piuttosto "semplice" o "a bassa dimensionalità". Vale a dire, è parametrizzato da poche quantità di interesse e la quantità di dati accessibili è grande. Ma nell'era dei **big-data**, il moderno signal processing e le attività di machine learning richiedono l'impiego dell'inferenza nel cosiddetto regime di **alta dimensionalità (alta-d)**. Ciò significa che anche se la quantità di dati fruibili è ingente, e la loro dimensionalità (molto) grande, anche il numero di parametri sconosciuti che caratterizzano il sistema in esame è enorme. Pertanto sono necessari strumenti statistici totalmente nuovi per dare un senso ai dati al fine di "estrarre il segnale dal rumore".

Statistica classica "a bassa dimensionalità"

Nella statistica classica il **segnale**, vale a dire l'informazione di interesse/il parametro sconosciuto da recuperare dai dati, è **a bassa dimensionalità**. Per essere più precisi, indichiamo il segnale $\mathbf{x} \in \mathbb{R}^p$ e i dati $\mathbf{y} = \mathbf{y}(\mathbf{x}) \in \mathbb{R}^n$, dipendenti dal segnale. Ad esempio, consideriamo un semplice esperimento in cui si cerca di dedurre se una moneta non è truccata, vale a dire se $\mathbb{P}(\text{testa}) = x = 1 - \mathbb{P}(\text{croce})$ con $x = 1/2$. In questo esempio lo spazio dei parametri ha dimensione $p = 1$ poiché il parametro rilevante -o segnale- $x \in [0, 1]$ è uno scalare. Un protocollo naturale per rispondere a questa domanda è: lanciamo la moneta n volte e registriamo il

language of statistical mechanics.

Statistical inference: old and new

Statistical inference aims at accurately describing some system, through the design of an appropriate probability distribution, based on data about this system and, potentially, some assumptions about it. Designing both statistically and computationally efficient inference procedures is thus crucial in virtually all fields of science.

Classical statistics is mostly concerned with the regime where the system of study is rather "simple", or "low-dimensional". Namely, it is parametrised by few quantities of interest and the amount of accessible data is large. But in the **big-data era**, contemporary signal processing and machine learning tasks require performing inference in the so-called **high-dimensional (high-d) regime**. This means that even if the amount of data as well at its dimensionality may be (very) large, the number of unknown parameters characterizing the system under study is also huge. Therefore totally new statistical tools are required to make sense of the data in order to "extract the signal from the noise".

Classical "low dimensional" statistics

In classical statistics the **signal**, namely the information of interest/unknown parameter to recover from the data, is **low-dimensional**. To be more precise, let us denote the signal $\mathbf{x} \in \mathbb{R}^p$ and the data $\mathbf{y} = \mathbf{y}(\mathbf{x}) \in \mathbb{R}^n$, that depends on the signal. For example consider a simple experiment where one tries to infer if a coin is fair, namely, whether $\mathbb{P}(\text{head}) = x = 1 - \mathbb{P}(\text{tail})$ with $x = 1/2$. In this example the parameter space has dimension $p = 1$ as the relevant parameter/signal $x \in [0, 1]$ is a scalar. A natural protocol to answer that question is: toss the coin n times and record the number $n_{\text{h}} \in \{0, \dots, n\}$ of times it fell on head. Then in the limit $n \gg 1$ the **law or large numbers**, which is a fundamental statistical property

numero $n_{\mathrm{h}} \in \{0, \dots, n\}$ di volte in cui è caduta sulla testa. Quindi nel limite $n \gg 1$ la **legge dei grandi numeri**, che è una proprietà statistica fondamentale al centro dell'apparente prevedibilità del mondo nonostante la sua intrinseca natura probabilistica, predice che la media empirica converge alla media statistica. Questo si traduce qui in $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\text{croce}_i = \text{testa}) = n_{\mathrm{h}}/n = x + o_n(1)$ con una correzione $o_n(1) \to 0$ come $n \to \infty$ (dove $\mathbb{1}(\cdot)$ è la funzione indicatore). Esempi meno basilari potrebbero essere inferire la costante gravitazionale terrestre dall'osservazione di $n \gg 1$ traiettorie di oggetti in caduta con varie condizioni iniziali (nel qual caso ancora $p = 1$), o inferire l'altezza media x_1, il peso medio x_2 e le varianze ad essi associate (x_3, x_4) sulla base di una vasta popolazione di n individui; in quest'ultimo caso $p = 4$.

Ciò che è veramente importante in questi esempi è che $p/n \ll 1$ è molto piccolo. In questo regime il modo ottimale per inferire i parametri dai dati, utilizzando un modello probabilistico per come i dati vengono generati condizionatamente ai parametri sconosciuti $\mathbf{x}$, è tramite la **stima di massima verosimiglianza (STV)**. Il modo probabilistico per rappresentare un processo casuale di generazione dei dati è una distribuzione di probabilità dell'osservazione dei dati condizionata ai parametri (sconosciuti) $\mathbb{P}(\mathbf{y} \mid \mathbf{x})$, chiamata **verosimiglianza**. Ad esempio, nell'esperimento del lancio della moneta, un'osservazione ovvia è che, in base al parametro di bias x, tutti i lanci sono indipendenti. Pertanto, mappando lo spazio binario dei dati $\{\text{testa}, \text{croce}\}$ in $\{0, 1\}$, ogni lancio risulta essere un esperimento di Bernoulli con $\mathbb{P}(y_i = 0 \mid x) = x$. Quindi la probabilità che la variabile casuale (v.c.) N_{h} assuma valore $k \in \{0, \dots, n\}$, dove N_{h} è il numero casuale di teste tra n prove, è la legge binomiale della probabilità di successo (sconosciuta) x: $\mathbb{P}(N_{\mathrm{h}} = k \mid x) = \binom{n}{k} x^k (1-x)^{nk}$. In questo esperimento gli unici dati sono il risultato n_{h} di N_{h} poiché l'ordine dei lanci è irrilevante. Pertanto $\mathbb{P}(N_{\mathrm{h}} = n_{\mathrm{h}} \mid x)$ è la probabilità dei dati.

È concettualmente utile introdurre la funzione di **verosimiglianza** $\mathcal{L}(\mathbf{x} \mid \mathbf{y}) := \mathbb{P}(\mathbf{y} \mid \mathbf{x})$. È importante pensare a $\mathcal{L}(\mathbf{x} \mid \mathbf{y})$ in realtà come una funzione dei parametri dati i dati (i dati sono fissi e non possono essere modificati), non il contrario come suggerito dalla distribuzione della proba-

at the core of the apparent predictability of the world in spite of its inherent probabilistic nature, predicts that the empirical mean converges to the statistical mean. This translates here to $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\text{toss}_i = \text{head}) = n_{\mathrm{h}}/n = x + o_n(1)$ with a correction $o_n(1) \to 0$ as $n \to \infty$ (here $\mathbb{1}(\cdot)$ is the indicator function). Less basic examples could be to infer the earth gravitational constant from the recording of $n \gg 1$ trajectories of falling objects with various initial conditons (in which case again $p = 1$), or infering the average height x_1, weight x_2 and the associated variances (x_3, x_4) based on some large population of n individuals; in the latter case $p = 4$.

What is really important in these examples is that $p/n \ll 1$ is very small. In this regime the optimal way to infer the parameters from the data, using a probabilistic model for how the data is generated conditional on the unknown paramaters $\mathbf{x}$, is through **maximum likelihood estimation (MLE)**. The probabilistic way to represent a random process of data generation is a probability distribution of observing the data conditional on the (unknown) parameters $\mathbb{P}(\mathbf{y} \mid \mathbf{x})$, called **likelihood**. For example, in the coin tossing experiment, an obvious observation is that conditional on the bias parameter x all tosses are independent. Therefore, mapping the binary data-space $\{\text{head}, \text{tail}\}$ to $\{0, 1\}$, each toss is a Bernoulli experiment with $\mathbb{P}(y_i = 0 \mid x) = x$. Therefore the likelihood that the random variable (r.v.) N_{h} takes value $k \in \{0, \dots, n\}$, N_{h} being the random number of heads among n trials, is the binomial law of (unknown) success probability x: $\mathbb{P}(N_{\mathrm{h}} = k \mid x) = \binom{n}{k} x^k (1-x)^{n-k}$. In this experiment the only data is the outcome n_{h} of N_{h} as the order of the tosses is irrelevant. Therefore $\mathbb{P}(N_{\mathrm{h}} = n_{\mathrm{h}} \mid x)$ is the likelihood of the data.

It is conceptually useful to introduce the **likelihood function** $\mathcal{L}(\mathbf{x} \mid \mathbf{y}) := \mathbb{P}(\mathbf{y} \mid \mathbf{x})$. It is important to think of $\mathcal{L}(\mathbf{x} \mid \mathbf{y})$ really as a function of the parameters given the data (the data being fixed and cannot be modified), not the other way around as suggested by the conditional proba-

bilità condizionata $\mathbb{P}(\mathbf{y} \mid \mathbf{x})$. Questa è la ragione dietro l'introduzione di una notazione specifica $\mathcal{L}(\mathbf{x} \mid \mathbf{y})$ che enfatizza questa corretta interpretazione. La SMV dice che si dovrebbe prendere come stima $\hat{\mathbf{x}} = \hat{\mathbf{x}}(\mathbf{y})$ dei parametri sconosciuti, essendo $\mathbf{x}$ il valore che massimizza il (logaritmo della) funzione di verosimiglianza in ragione dei dati osservati. Nel caso del lancio di una moneta, $\mathcal{L}(x \mid n_\mathrm{h}) := \mathbb{P}(N_\mathrm{h} = n_\mathrm{h} \mid x)$, quindi la SMV dà

$$\begin{aligned}\hat{x} &\in \operatorname*{argmax}_{x\in[0,1]} \ln \mathcal{L}(x \mid n_\mathrm{h}) \\ &= \operatorname*{argmax}_{x\in[0,1]} \{n_\mathrm{h} \ln x + (n - n_\mathrm{h}) \ln(1-x)\}.\end{aligned}$$

La funzione $\{\cdots\}$ è concava, quindi il suo unico massimizzatore è facilmente individuabile come $\hat{x}(n_\mathrm{h}) = n_\mathrm{h}/n$. L'approccio del principio della massima verosimiglianza consente quindi di recuperare la scelta naturale suggerita dalla legge dei grandi numeri. Questo metodo semplice ma potente ha guidato la statistica classica per più di un secolo dal suo sviluppo da parte di C. Gauss, P. S. Laplace o F. Edgeworth. Infine, si può dimostrare che la SMV è ottimale nel limite di $p/n \to 0$ in un senso preciso e abbastanza generale [1].

Statistica ad alta dimensionalità

Il regime ad alta-d si riferisce generalmente ad approcci statistici nei quali sia il numero di parametri che la popolosità dei dati, che possono essere essi stessi ad alta-d, sono ampi e comparabili: $p/n \to \delta > 0$ ed entrambi $p, n \to \infty$ in maniera tale che δ risulti una costante di ordine uno. Questo è in contrasto con il limite classico $\delta \to 0$. Per essere più precisi, abbiamo bisogno di $p/(n \times \mathrm{RSR_d}) \to \delta > 0$, dove il **rapporto segnale / rumore (RSR)**, per dato $\mathrm{RSR_d}$, è una misura del contenuto informativo di $\mathbf{x}$ trasportato da un singolo punto dei dati, in media. La definizione di $\mathrm{RSR_d}$ dipende dal modello in esame, ma è sempre correlata ad un modo naturale di confron-

[1] Lo stimatore di SMV $\hat{\mathbf{x}}$ è ottimale nel seguente senso: nell'impostazione bayesiana ottimale che sarà descritta in seguito, nel regime $n \gg p$, $\hat{\mathbf{x}}$ è sia uno stimatore di Bayes (cioè minimizza il rischio di Bayes associato alla distribuzione del segnale $\mathbb{P}(\mathbf{x})$) sia un minimax per la perdita quadratica media: si veda la prossima sezione sulla teoria delle decisioni ed il Capitolo 12 del celebre libro [1].

bility distribution $\mathbb{P}(\mathbf{y} \mid \mathbf{x})$. This is the reason behind the introduction of a specific notation $\mathcal{L}(\mathbf{x} \mid \mathbf{y})$ emphasising this correct interpretation. MLE says that one should take as estimate $\hat{\mathbf{x}} = \hat{\mathbf{x}}(\mathbf{y})$ of the unknown parameters $\mathbf{x}$ the value that maximizes the (logarithm of the) likelihood function given the measured data. In the coin tossing case, $\mathcal{L}(x \mid n_\mathrm{h}) := \mathbb{P}(N_\mathrm{h} = n_\mathrm{h} \mid x)$, so MLE gives

$$\begin{aligned}\hat{x} &\in \operatorname*{argmax}_{x\in[0,1]} \ln \mathcal{L}(x \mid n_\mathrm{h}) \\ &= \operatorname*{argmax}_{x\in[0,1]} \{n_\mathrm{h} \ln x + (n - n_\mathrm{h}) \ln(1-x)\}.\end{aligned}$$

The function $\{\cdots\}$ is concave so its unique maximiser is easily found to be $\hat{x}(n_\mathrm{h}) = n_\mathrm{h}/n$. The principled approach of MLE therefore allows to recover the natural choice suggested by the law of large numbers. This simple but powerful method has driven classical statistics for more than a century since its development by C. Gauss, P. S. Laplace or F. Edgeworth. Finally, MLE can be shown to be optimal in the limit $p/n \to 0$ in a precise and quite general sense[1].

High-dimensional statistics

The high-d regime generally refers to statistical settings in which both the number of parameters and of data points –that can themselves be high-d– are large and comparable: $p/n \to \delta > 0$ as both $p, n \to \infty$ together, with δ an order one constant. This is to be contrasted with the classical limit $\delta \to 0$. To be more precise we require $p/(n \times \mathrm{SNR_d}) \to \delta > 0$, where the **signal-to-noise ratio (SNR)** per data point $\mathrm{SNR_d}$ is a measure of the information content about $\mathbf{x}$ carried by a single data point, in average. The definition of $\mathrm{SNR_d}$ depends on the model under study, but is always related to a natural way of comparing the (average) signal amplitude with the one of the noise that corrupts the data. When it is of or-

[1] The MLE estimator $\hat{\mathbf{x}}$ is optimal in the following sense: in the Bayesian optimal setting described soon, in the regime $n \gg p$, $\hat{\mathbf{x}}$ is both a Bayes estimator (namely, it minimizes the Bayes risk associated with the distribution of the signal $\mathbb{P}(\mathbf{x})$) and minimax for the mean-square loss, see the upcoming section on decision theory and Chapter 12 in the great book [1].

tare l'ampiezza (media) del segnale con quella del rumore che corrompe i dati. Quando questo rapporto è di ordine uno, si recupera la solita definizione di regime ad alta-d $p/n \to \delta > 0$.

Nell'esperimento del lancio di monete, un modo naturale per quantificare il rumore che corrompe il singolo punto dei dati $y = n_{\mathrm{h}} \in \{0, \dots, n\}$ è dato dalla varianza del valore associato alla v.c. N_{h}. La varianza della distribuzione binomiale $\mathrm{Bin}(n, x)$ è $nx(1-x)$ quindi $\mathrm{RSR_d} = nx(1-x)$ è grande. Quindi $\#\text{parametri} \div (\#\text{dati} \times \mathrm{RSR_d}) = 1/(1 \times nx(1-x)) \to 0$ quando $n \to \infty$. Un'interpretazione altrettanto valida è: abbiamo n punti dati $(y_i)_{i=1}^n$, ciascuno dei quali è il risultato di un esperimento di Bernoulli. Ogni $y_i \in \{0, 1\}$ ha varianza $x(1-x) = O(1)$. Quindi $\#\text{parametri} \div (\#\text{dati} \times \mathrm{RSR_d}) = 1/(n \times x(1-x)) \to 0$. Poiché tende a 0 siamo nel regime classico della statistica.

Il regime della statistica in esame

$$p/(n \times \mathrm{RSR_d}) \to \delta > 0$$

è particolarmente rilevante per le applicazioni in tutti i tipi di compiti di elaborazione del segnale (elaborazione di immagini e suoni, applicazioni biomedicali, codici per la correzione degli errori per le comunicazioni, etc.) e nell'apprendimento automatico (classificazione automatica delle immagini, scoperte di farmaci, elaborazione e traduzione del linguaggio naturale, auto a guida autonoma, etc.) che stanno cambiando il mondo con una pasta senza precedenti. Un esempio sono le moderne **reti neurali profonde** addestrate su basi di dati con milioni di immagini. Ma il numero di parametri (i.e. pesi sinaptici) che definiscono questi modelli complessi è dello stesso ordine, o anche molto più grande.

Non è esagerato chiamarlo **data-revolution**, e la statistica ad alta-d è il suo core teorico. Gli altri pilastri di questa rivoluzione sono la quantità di dati accessibili, nonché i computer moderni, con unità di calcolo specifiche in grado di elaborare set di dati così enormi, quali le unità di elaborazione grafica (GPU).

La comprensione del regime ad alta-d richiede concetti e strumenti matematici totalmente nuovi, e per risolvere i problemi applicati reali, abbiamo bisogno di nuovi algoritmi. Parlando di algoritmi, nel mondo ad alta-d in cui i dati sono massivi,

der one we recover the usual definition of high-d regime $p/n \to \delta > 0$.

In the coin tossing experiment, a natural way of quantifying the noise corrupting the single data point $y = n_{\mathrm{h}} \in \{0, \dots, n\}$ is given by the variance of the associated r.v. N_{h}. The variance of the binomial distribution $\mathrm{Bin}(n, x)$ is $nx(1-x)$ so $\mathrm{SNR_d} = nx(1-x)$ is large. Therefore $\#\text{paramaters} \div (\#\text{data points} \times \mathrm{SNR_d}) = 1/(1 \times nx(1-x)) \to 0$ as $n \to \infty$. An equally valid interpretation is: we have n data points $(y_i)_{i=1}^n$, each one being the outcome of a Bernoulli experiment. Each $y_i \in \{0, 1\}$ has variance $x(1-x) = O(1)$. Then $\#\text{paramaters} \div (\#\text{data points} \times \mathrm{SNR_d}) = 1/(n \times x(1-x)) \to 0$. Because it tends to 0 we are in the classical regime of statistics.

The modern statistical regime

$$p/(n \times \mathrm{SNR_d}) \to \delta > 0$$

is particularly relevant for applications in all sorts of signal processing tasks –image and sound processing, medical applications, error-correcting codes for communications, etc– and machine learning –automatic image classification, drug discoveries, natural language processsing and translation, self-driving cars, etc– that are changing the world at a unprecedented paste. One example are the modern **deep neural networks** trained on data-bases with millions images. But the number of parameters/synaptic weights defining these complex models is of the same order, or even much bigger.

It is not exagarating to call that a **data-revolution**, and high-d statistics is the theoretical powerhouse at its core. The other key pillars of this revolution being the amount of accessible data, as well as the modern computers and specific computational units able to process such huge data sets, like graphical processing units (GPUs).

Understanding the high-d regime requires totally new concepts and mathematical tools, and for solving actual applied problems, we need new algorithms. Speaking about algorithms, in the high-d world where the data is so massive,

no così massicci, non solo l'**efficienza statistica** è importante –a dire la capacità di un algoritmo di estrarre le informazioni rilevanti, indipendentemente da qualsiasi "problema di velocità"– ma anche l'**efficienza computazionale**, poiché questa può repentinamente diventare un collo di bottiglia. I ricercatori che lavorano in applicazioni di statistica ad alta-d devono sempre tenere presente queste due considerazioni, una caratteristica fondamentale del campo che lo rende così interessante e stimolante allo stesso tempo.

In questo articolo focalizzeremo principalmente la nostra attenzione sulle limitazioni dell'inferenza da una prospettiva di **teoria dell'informazione** (o **statistica**), tralasciando le considerazioni algoritmiche.

not only **statistical efficiency** matters –namely, the capacity of an algorithm to extract the relevant information, independently of any "speed concern"–, but also **computational efficiency**, as it quickly becomes a bottleneck. Researchers working in applications of high-d statistics must always keep in mind these two considerations, a key feature of the field that makes it so interesting and challenging at the same time.

In this article we will mainly focus our attention on the **information-theoretic** (or **statistical**) limitations to inference and leave the algorithmic considerations aside.

Nozioni di base sull'inferenza bayesiana

L'uso della conoscenza a-priori per allegerire il fardello della dimensionalità

Abbiamo detto che nella statistica classica la stima SMV è ottimale. Questo non è più vero nel regime ad alta-d. Poiché la quantità di dati è paragonabile al numero di parametri sconosciuti da dedurre, ciò potrebbe creare degenerazioni nella soluzione di SMV, nel senso che l'insieme $\operatorname{argmax}_{\mathbf{x}\in\mathbb{R}^p} \ln \mathcal{L}(\mathbf{x} \mid \mathbf{y})$ può avere un enorme cardinalità ove tutte le soluzioni sono ugualmente pessime. Questo è un problema relativo al **fardello della dimensionalità**: nel regime ad alta-d il volume dello spazio in cui vive il segnale $\mathbf{x}$ aumenta così velocemente (esponenzialmente veloce con p) che i dati disponibili restano sempre relativamente scarsi. Questa scarsità è problematica per qualsiasi metodo che richieda significatività statistica. Per ottenere un risultato statisticamente valido e affidabile, la quantità di dati necessari per supportare il risultato spesso cresce in modo esponenziale con la dimensionalità, e tale volume non è mai accessibile esaustivamente (si provi a calcolare $\exp(p)$ con $p = 10, 100, 1000...$).

Quindi si deve colmare questa "lacuna informativa" a causa della relativa mancanza di dati usando **ipotesi** su $\mathbf{x}$, ovvero **conoscenza a-priori**. Citando D. Mackay: "you cannot do inference without making assumptions", si veda lo splendido libro [2]. Ciò può essere formalizzato attra-

Basics of Bayesian inference

Breaking the curse of dimensionality using a-priori knowledge

We have said that in classical statistics MLE estimation is optimal. This is not true anymore in the high-d regime. Because the amount of data is comparable to the number of unknown parameters to infer, this may create degeneracies in the solution of MLE, in the sense that the set $\operatorname{argmax}_{\mathbf{x}\in\mathbb{R}^p} \ln \mathcal{L}(\mathbf{x} \mid \mathbf{y})$ may have a huge cardinal, all solutions being equally bad. This is one issue related to the **curse of dimensionality**: in the high-d regime the volume of the space in which lives the signal $\mathbf{x}$ increases so fast (exponentially fast with p) that the available data becomes relatively sparse. This sparsity is problematic for any method that requires statistical significance. In order to obtain a statistically sound and reliable result, the amount of data needed to support the result often grows exponentially with the dimensionality, but such an amount is never accessible (compute $\exp p$ with $p = 10, 100, 1000...$).

Therefore one has to fill-in this "information gap" due to relative lack of data using **assumptions** about $\mathbf{x}$, namely, **a-priori knowledge**. Quoting D. Mackay: "you cannot do inference without making assumptions", see the amazing book [2]. This can be formalised through a probabil-

verso una distribuzione di probabilità $\mathbb{P}(\mathbf{x})$ che dipende solo dal segnale, e quindi è completamente indipendente dai dati $\mathbf{y}$. Questa distribuzione, chiamata **prior**, traduce nel linguaggio della probabilità l'intero insieme di assunzioni a priori fatte dallo statistico circa $\mathbf{x}$, prima che i dati vengano raccolti. È fondamentale che una volta acquisiti i dati, la prior non venga modificata di conseguenza, altrimenti ciò potrebbe creare una distorsione interpretativa (i.e. **bias**). Tali ipotesi potrebbero essere, ad esempio, che il segnale sia binario $\mathbf{x} \in \{-1, 1\}^p$ con componenti (i.i.d.) indipendenti e identicamente distribuite uniformemente x_i (come i bit ricevuti da alcune sorgenti di comunicazioni). Questa assunzione di base si tradurrebbe in $\mathbb{P}(\mathbf{x}) = \prod_{i=1}^{p} \frac{1}{2}(\delta_{x_i,-1} + \delta_{x_i,1})$. Ma forse in aggiunta si può sapere che in realtà il segnale è **sparso**[2], il che significa che ha un frazione ρ di ingressi pari a 0. In questo caso $\mathbb{P}(\mathbf{x}) = \prod_{i=1}^{p}\{\rho\delta_{x_i,0} + \frac{1-\rho}{2}(\delta_{x_i,-1} + \delta_{x_i,1})\}$. E così via: più ricca è la serie di ipotesi, più complessa è l'espressione precedente.

Combinazione di ipotesi e dati: la formula di Bayes

Una delle più eleganti affermazioni probabilistiche in assoluto è la cosiddetta **formula di Bayes**. È stata scoperta da un reverendo di nome T. Bayes prima della sua morte nel 1761. Indipendentemente da Bayes, P. S. Laplace formalizzò idee simili nel 1774. Jeffreys, uno dei padri delle moderne statistiche bayesiane, scrisse in seguito che "questo teorema sta alla teoria della probabilità come il teorema di Pitagora sta alla geometria". Per quanto semplice, racchiude in un'unica equazione una potente ricetta, più utile che mai nel contesto dell'inferenza in alta-d. Questa afferma infatti che il modo corretto per combinare conoscenza a priori e dati è attraverso una semplice

[2] L'assunzione di scarsità permette di ricostruire segnali ad altissima dimensionalità da relativamente pochi punti dati, e, ad esempio, di invertire sistemi apparentemente sottodeterminati di lineari equazioni. Questo costituisce la base di un intero campo di ricerca in matematica ed, in particolare, nell'elaborazione del segnale, chiamato **compressive sensing**, si veda l'eccellente introduzione [3] o [23] per un approccio di meccanica statistica.

ity distribution $\mathbb{P}(\mathbf{x})$ that depends on the signal only, and therefore is completely independent of the data $\mathbf{y}$. This distribution, called **prior**, translates in the language of probability the whole set of a-priori assumptions made by the statistician about $\mathbf{x}$, before that the data is collected. It is crucial that once the data is acquired, the prior is not modified accordingly, otherwise this may create **bias**. Such assumptions could be, for example, that the signal is binary $\mathbf{x} \in \{-1, 1\}^p$ with uniform independent and identically distributed (i.i.d.) components x_i (like the bits received from some communication source). This very basic assumption would translate in $\mathbb{P}(\mathbf{x}) = \prod_{i=1}^{p} \frac{1}{2}(\delta_{x_i,-1} + \delta_{x_i,1})$. But maybe in addition one may know that actually the signal is **sparse**[2], meaning it has a fraction ρ of entries equal to 0. In this case $\mathbb{P}(\mathbf{x}) = \prod_{i=1}^{p}\{\rho\delta_{x_i,0} + \frac{1-\rho}{2}(\delta_{x_i,-1} + \delta_{x_i,1})\}$. And so on: the richer the set of assumptions, the more complex is the prior expression.

Combining assumptions and data: Bayes formula

One of the most elegant probabilistic statement ever is the so-called **Bayes formula**. It has been discovered by a reverend named T. Bayes before his death in 1761. Independently of Bayes, P. S. Laplace formalised similar ideas in 1774. Jeffreys, one of the father of modern Bayesian statistics, later wrote that Bayes's theorem "is to the theory of probability what the Pythagorean theorem is to geometry". As simple as it is, it encapsulates in a single equation a powerful recipe, more useful than ever in the context of inference in high dimensions. It says that the proper way to combine a-priori knowledge and data is through a simple multiplication followed by a normalization:

$$\mathbb{P}(\mathbf{x} \mid \mathbf{y}) = \frac{\mathbb{P}(\mathbf{x})\mathbb{P}(\mathbf{y} \mid \mathbf{x})}{\mathbb{P}(\mathbf{y})}. \qquad (1)$$

[2] The sparsity assumption allows to reconstruct very high-d signals from relatively few data points, and, e.g., to invert apparently under-determined systems of linear equations. This forms the basis of a whole field of research in mathematics and signal processing called **compressive sensing**, see the excellent introduction [3] or [23] for a statistical mechanics approach.

moltiplicazione seguita da una normalizzazione:

$$\mathbb{P}(\mathbf{x} \mid \mathbf{y}) = \frac{\mathbb{P}(\mathbf{x})\mathbb{P}(\mathbf{y} \mid \mathbf{x})}{\mathbb{P}(\mathbf{y})}. \tag{1}$$

Cioè

posterior $=$ (prior $\times$ verosimiglianza)/evidenza

La distribuzione a posteriori, i.e. **posterior** $\mathbb{P}(\mathbf{x} \mid \mathbf{y})$ significa $\mathbb{P}$ (i parametri hanno valore $\mathbf{x}$ assunto che i dati sono $\mathbf{y}$). "Posterior" è nel senso "dopo che i dati sono stati raccolti". La distribuzione a posteriori è quindi la moltiplicazione della prior, che formalizza tutte le nostre ipotesi sul segnale, con la verosimiglianza, che modella il processo di generazione dei dati condizionato al segnale. La posterior combina in un'unica distribuzione di probabilità tutte le informazioni che abbiamo sul segnale, così come la nostra incertezza su di esso; ad esempio, la varianza della posterior $\mathrm{Var}(\mathbf{x} \mid \mathbf{y}) = \mathbb{E}[\|\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{y}]\|^2 \mid \mathbf{y}]$ (definendo $\mathbb{E}[g(\mathbf{x}) \mid \mathbf{y}] := \int d\mathbf{x}\, g(\mathbf{x})\, \mathbb{P}(\mathbf{x} \mid \mathbf{y})$).

La normalizzazione $\mathbb{P}(\mathbf{y}) = \int d\mathbf{x}'\, \mathbb{P}(\mathbf{x}')\mathbb{P}(\mathbf{y} \mid \mathbf{x}')$ si chiama **evidenza**: è la distribuzione marginale dei dati. Si noti che in dimensioni elevate questa distribuzione può essere molto difficile, se non impossibile, da calcolare esattamente poiché richiede di eseguire un integrale p dimensionale, con p molto grande.

Limiti teorici dell'informazione: l'impostazione bayesiana ottimale

D'ora in poi limiteremo la nostra discussione all'**impostazione bayesiana ottimale**. Ciò significa che lo statistico conosce il modello alla base del processo di generazione dei dati (ma ovviamente non il segnale $\mathbf{x}$), che si traduce nella corretta probabilità $\mathbb{P}(\mathbf{y} \mid \mathbf{x})$. Inoltre sfrutta correttamente tutte le informazioni a priori sul segnale, vale a dire, il segnale è stato effettivamente generato casualmente dalla prior $\mathbb{P}(\mathbf{x})$ utilizzata dallo statistico. Pertanto, in questo contesto, la distribuzione a posteriori è quella "corretta" e tutta la discussione imminente si applica solo a questo caso.

L'impostazione bayesiana ottimale è fondamentale: come diremo, qualsiasi **stimatore ottimale**, qualunque nozione significativa di ottimalità sia considerata, si basa su una corretta posterior. Per-

That is,

posterior $=$ prior $\times$ likelihood/evidence.

The **posterior distribution** $\mathbb{P}(\mathbf{x} \mid \mathbf{y})$ signifies $\mathbb{P}($ the parameters take value $\mathbf{x}$ given the data is $\mathbf{y}$ and our a-priori knowledge). "Posterior" is in the sense of a-posteriori that the data has been collected. The posterior distribution is therefore the multiplication of the prior, that formalises all our assumptions about the signal, with the likelihood, that models the data-generating process conditional on the signal. The posterior combines in a single probability distribution all information we have about the signal, as well as our uncertainty about it through, e.g., the posterior variance $\mathrm{Var}(\mathbf{x} \mid \mathbf{y}) = \mathbb{E}[\|\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{y}]\|^2 \mid \mathbf{y}]$ (defining $\mathbb{E}[g(\mathbf{x}) \mid \mathbf{y}] := \int d\mathbf{x}\, g(\mathbf{x})\, \mathbb{P}(\mathbf{x} \mid \mathbf{y})$).

The normalization $\mathbb{P}(\mathbf{y}) = \int d\mathbf{x}'\, \mathbb{P}(\mathbf{x}')\mathbb{P}(\mathbf{y} \mid \mathbf{x}')$ is called **evidence**. It is the marginal distribution of the data. Note that in high dimensions this distribution can be very hard, if not impossible, to compute exactly as it requires to perform a p-dimensional integral, with p very large.

Information-theoretic limits: the Bayesian optimal setting

From now on we will restrict our discussion to the **Bayesian optimal setting**. This means that the statistician knows the model underlying the data-generating process (but of course not the signal $\mathbf{x}$), which translates into the correct likelihood $\mathbb{P}(\mathbf{y} \mid \mathbf{x})$. In addition she also exploits correctly all a-priori information about the signal, namely, the signal has indeed been randomly generated from the prior $\mathbb{P}(\mathbf{x})$ used by her. Therefore in this setting, the posterior distribution is the "correct one" and all the upcoming discussion applies only to this case.

The Bayesian optimal setting is fundamental: as we will argue, any **optimal estimator**, whatever meaningful notion of optimality is considered, relies on the correct posterior. Therefore in order to study the **information-theoretical**

tanto, per studiare i **limiti teorici dell'informazione** nell'inferenza -vale a dire i migliori risultati a cui si può ambire indipendentemente da qualsiasi preoccupazione computazionale-, è necessario studiare l'inferenza proprio in questo contesto ottimale. Le prestazioni degli stimatori per $\mathbf{x}$ in questa impostazione non possono essere superate da alcun algoritmo, nemmeno da quelli che possono essere eseguiti per un tempo infinito. L'impostazione bayesiana ottimale è al centro della **teoria dell'informazione** [14].

Un'impostazione diversa in cui la probabilità utilizzata non descrive correttamente la generazione dei dati e/o la prior è distorta (cioè non corrisponde alla distribuzione di probabilità da cui è stato generato il segnale) è molto più complicata e va oltre la presente discussione. Nell'esperimento del lancio di monete ripetuto, un'ipotesi a priori errata potrebbe essere che i lanci siano indipendenti, mentre per qualche motivo i lanci sono correlati; questo potrebbe essere dovuto, ad esempio, a un minuscolo demone che vive all'interno della moneta e ciò modificherebbe la probabilità $x = x_i(y_{i-1})$ di testa per il lancio i in funzione del risultato del lancio precedente y_i. Tuttavia, gran parte della fenomenologia che è inerente all'alta dimensionalità, rimane la stessa nelle impostazioni bayesiane ottimali e nelle (più realistiche) non-ottimali.

Ottimalità degli stimatori: teoria Bayesiana della decisione

Per quantificare le prestazioni dello statistico nella stima di $\mathbf{x}$, dobbiamo definire un corretto errore di ricostruzione associato a un dato stimatore $\hat{\mathbf{x}} = \hat{\mathbf{x}}(\mathbf{y})$. Questa metrica di errore è chiamata **loss function** nelle statistica e può essere pensata come una funzione energia. Ci sono molte scelte possibili, la cui rilevanza dipende dall'applicazione specifica. Una scelta canonica è la loss function $0-1$: $\ell(\hat{\mathbf{x}}, \mathbf{x}) = 1 - \mathbb{1}(\hat{\mathbf{x}} = \mathbf{x})$. Questa scelta ha senso quando il segnale è discreto, come nelle comunicazioni in cui il segnale è composto da bits. La loss function non può essere calcolata poiché dipende dal segnale sconosciuto. Quindi per definire una nozione di "bontà" di uno stimatore definiamo il **rischio a posteriori** (la nostra

limits of inference –namely the best results one can aim for independently of any computational concern–, one needs to precisely study inference in this optimal setting. The performance of estimators for $\mathbf{x}$ in this setting cannot be outperformed by any algorithm, even those allowed to run for infinite time. The Bayesian optimal setting is at the core of **information theory** [14].

The mismatched setting where the likelihood used is not properly describing the data generation and/or the prior is biased (i.e., does not correspond to the probability distribution from which the signal was generated) is much more complicated and goes beyond the present discussion. In the repeated coin tossing experiment, a wrong a-priori assumption could be that the tosses are independent, while for some reason the tosses were correlated; this could be due, e.g., to a tiny demon living inside the coin and that would modify the probability $x = x_i(y_{i-1})$ of head for the toss i as a function of the previous tossing result y_i. Yet, a lot of the phenomenology, that is inherent to the high-dimensionality, remains the same in the Bayesian optimal and mismatched (more realistic) settings.

Optimality of estimators: Bayesian decision theory

In order to quantify the performance of the statistician in estimating $\mathbf{x}$, we need to define a proper recontruction error associated with a given estimator $\hat{\mathbf{x}} = \hat{\mathbf{x}}(\mathbf{y})$. This error metric is called **loss** in statistics, and can be thought as an energy function. There are many possible choices, whose relevance depends on the specific application at hand. One canonical choice is the $0-1$ loss: $\ell(\hat{\mathbf{x}}, \mathbf{x}) = 1 - \mathbb{1}(\hat{\mathbf{x}} = \mathbf{x})$. This choice makes sense when the signal is discrete, like in communications where the signal is made of bits. The loss cannot be computed as it depends on the unknown signal. Therefore in order to define a notion of "goodness" of an estimator we define the **posterior risk** (our best estimate of the loss):

$$r(\hat{\mathbf{x}} \mid \mathbf{y}) := \int d\mathbf{x}\, \mathbb{P}(\mathbf{x} \mid \mathbf{y}) \ell(\hat{\mathbf{x}}, \mathbf{x}),$$

migliore stima della perdita):

$$r(\hat{\mathbf{x}} \mid \mathbf{y}) := \int d\mathbf{x}\, \mathbb{P}(\mathbf{x} \mid \mathbf{y}) \ell(\hat{\mathbf{x}}, \mathbf{x}),$$

che è semplicemente uguale a $1 - \mathbb{P}(\hat{\mathbf{x}} \mid \mathbf{y})$ per la loss function di $0-1$, o la sua media rispetto all'evidenza, chiamata **rischio di Bayes** $r(\hat{\mathbf{x}}) := \int d\mathbf{y}\, \mathbb{P}(\mathbf{y}) r(\hat{\mathbf{x}} \mid \mathbf{y})$. Entrambi possono essere calcolati in teoria solo dalla conoscenza dei dati e del modello statistico sottostante; in pratica questo può essere computazionalmente molto impegnativo ad alta dimensionalità.

Con un conto diretto si vede che lo stimatore ottimale, ottimale nel senso di minimizzare il rischio a posteriori (o di Bayes), è nel caso di loss function $0-1$ dato dalla moda a posteriori:

$$\hat{\mathbf{x}}_{\mathrm{MAP}}(\mathbf{y}) := \operatorname*{argmin}_{\hat{\mathbf{x}} \in \mathbb{R}^p} r(\hat{\mathbf{x}} \mid \mathbf{y}) = \operatorname*{argmax}_{\hat{\mathbf{x}} \in \mathbb{R}^p} \mathbb{P}(\hat{\mathbf{x}} \mid \mathbf{y}).$$

MAP sta per stimatore **massimo a posteriori**. Un'altra scelta comune più appropriata per i segnali a valori reali è la loss function di L_2: $\ell(\hat{\mathbf{x}}, \mathbf{x}) = \|\hat{\mathbf{x}} - \mathbf{x}\|^2$. L'associato rischio a posteriori è chiamato **errore quadratico medio**, e lo stimatore che lo minimizza è lo stimatore **errore quadratico medio minimo (MMSE)**:

$$\begin{aligned} \hat{\mathbf{x}}_{\mathrm{MMSE}}(\mathbf{y}) &:= \operatorname*{argmin}_{\hat{\mathbf{x}} \in \mathbb{R}^p} \int d\mathbf{x}\, \mathbb{P}(\mathbf{x} \mid \mathbf{y}) \|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \\ &= \int d\mathbf{x}\, \mathbf{x}\, \mathbb{P}(\mathbf{x} \mid \mathbf{y}) =: \mathbb{E}[\mathbf{x} \mid \mathbf{y}]. \quad (2) \end{aligned}$$

La seconda uguaglianza è facilmente dimostrata equiparando il gradiente rispetto a $\hat{\mathbf{x}}$ del rischio a posteriori al vettore di tutti zeri (e mediante convessità si arriva all'unico minimizzatore). Pertanto lo stimatore MMSE è "semplicemente" la media a posteriori. Ovviamente in generale questo può essere molto costoso da calcolare perché ci sono due integrali p dimensionali: l'evidenza (necessaria per normalizzare la posterior), e quindi l'integrale $\int d\mathbf{x}\, \mathbf{x}\, \mathbb{P}(\mathbf{x} \mid \mathbf{y})$. Pertanto in molte applicazioni pratiche si preferisce lo stimatore MAP poiché aggira questo problema (ma non è ottimale per qualsiasi altra loss function rispetto alla loss function $0-1$). Il principio che per una data loss function/rischio (naturale) lo stimatore ottimale associato si basa sulla posterior è generale.

Concentrandosi sulla loss function in L_2, l'errore di inferenza associato allo stimatore MMSE è, naturalmente, il **minimo errore quadratico**

which is simply equal to $1 - \mathbb{P}(\hat{\mathbf{x}} \mid \mathbf{y})$ for the $0-1$ loss, or its average with respect to the evidence, called **Bayes risk** $r(\hat{\mathbf{x}}) := \int d\mathbf{y}\, \mathbb{P}(\mathbf{y}) r(\hat{\mathbf{x}} \mid \mathbf{y})$. Both can be in theory computed from the knowledge of the data only and the underlying statistical model; in practice this may be computationally very demanding in high dimension.

We directly get that the optimal estimator, optimal in the sense of minimizing the posterior (or Bayes) risk, is in the case of $0-1$ loss given by the posterior mode:

$$\hat{\mathbf{x}}_{\mathrm{MAP}}(\mathbf{y}) := \operatorname*{argmin}_{\hat{\mathbf{x}} \in \mathbb{R}^p} r(\hat{\mathbf{x}} \mid \mathbf{y}) = \operatorname*{argmax}_{\hat{\mathbf{x}} \in \mathbb{R}^p} \mathbb{P}(\hat{\mathbf{x}} \mid \mathbf{y}).$$

MAP stands for **maximum a-posteriori** estimator. Another common choice that is more appropriate for real-valued signals is the L_2 loss: $\ell(\hat{\mathbf{x}}, \mathbf{x}) = \|\hat{\mathbf{x}} - \mathbf{x}\|^2$. The associated posterior risk is called the **mean-square error**, and the estimator that minimizes it is the **minimum mean-square error (MMSE) estimator**:

$$\begin{aligned} \hat{\mathbf{x}}_{\mathrm{MMSE}}(\mathbf{y}) &:= \operatorname*{argmin}_{\hat{\mathbf{x}} \in \mathbb{R}^p} \int d\mathbf{x}\, \mathbb{P}(\mathbf{x} \mid \mathbf{y}) \|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \\ &= \int d\mathbf{x}\, \mathbf{x}\, \mathbb{P}(\mathbf{x} \mid \mathbf{y}) =: \mathbb{E}[\mathbf{x} \mid \mathbf{y}]. \quad (2) \end{aligned}$$

The second equality is easily shown by equating the gradient w.r.t. $\hat{\mathbf{x}}$ of the posterior risk to the all-zeros vector (by convexity it leads the unique minimizer). Therefore the MMSE estimator is "simply" the posterior mean. Of course in general this may be very costly to compute because there are two p-dimensional integrals: the evidence (necessary to normalize the posterior), and then the integral $\int d\mathbf{x}\, \mathbf{x}\, \mathbb{P}(\mathbf{x} \mid \mathbf{y})$. Therefore in many practical applications one prefers the MAP estimator as it bypasses this issue (but is sub-optimal for any other loss than the $0-1$ loss). The principle that for a given (natural) loss/risk the associated optimal estimator relies on the posterior is general.

Focusing on the L_2 loss, the inference error associated with the MMSE estimator is, naturally,

medio:

$$\text{MMSE}_p := \tfrac{1}{p}\|\mathbb{E}[\mathbf{x} \mid \mathbf{y}] - \mathbf{x}\|^2,$$
$$\text{MMSE} := \lim_p \mathbb{E}_{\mathbf{x},\mathbf{y}} \text{MMSE}_p. \qquad (3)$$

Per $\lim_p$ si intende sempre il " limite termodinamico " $p \to \infty$. Consideriamo l'errore medio: il simbolo $\mathbb{E}$ significa l'attesa rispetto al segnale $\mathbf{x}$ e al dato $\mathbf{y}$ (condizionato a $\mathbf{x}$), visti come v.c. estratte dalle rispettive distribuzioni (a priori e di probabilità). Questa quantità riassume in un unico numero tutta la complessità del problema ad alta-d, fornendo l'errore ottimale a cui si può ambire per qualsiasi algoritmo, in media su tutte le possibili realizzazioni del problema.

Ci si potrebbe chiedere se il numero MMSE sia sufficiente per descrivere il problema, poiché, a priori, potrebbe verificarsi anche il caso che la v.c. MMSE_p (casuale per via di $(\mathbf{x}, \mathbf{y})$) fluttui molto (cioè, abbia varianza $O(1)$). Ma in problemi di inferenza ad alta-d ben definiti, nell'impostazione ottimale bayesiana questo **si concentra**; si dice che sia **automediante** nella terminologia fisica. Ciò significa che in realtà non oscilla molto per i sistemi di grandi dimensioni e diventa deterministico nel limite $p, n \to \infty$:

$$\text{MMSE}_p = \text{MMSE} + o_p(1)$$

dove $o_p(1)$ è per definizione una quantità tale che $\lim_p o_p(1) = 0$. Pertanto è sufficiente concentrarsi sull'MMSE medio asintotico per prevedere il comportamento di istanze fisse (tipiche) del problema. L'auto-media delle metriche di errore nell'inferenza bayesiana ottimale ad alta-d è molto generica [4, 5] (cosa non necessariamente vera in scenari non-ottimali): è una manifestazione non banale del fenomeno della **concentrazione in misura** nei modelli ad alta-d, che è al centro del determinismo/prevedibilità di questi complessi sistemi casuali.

Inferenza ad alta dimensionalità come meccanica statistica

Stabiliamo ora delle chiare connessioni tra quello che abbiamo discusso fino ad ora inerentemente l'inferenza ad alta-d e la meccanica statistica.

the **miminum mean-square error**:

$$\text{MMSE}_p := \tfrac{1}{p}\|\mathbb{E}[\mathbf{x} \mid \mathbf{y}] - \mathbf{x}\|^2,$$
$$\text{MMSE} := \lim_p \mathbb{E}_{\mathbf{x},\mathbf{y}} \text{MMSE}_p. \qquad (3)$$

By $\lim_p$ we always mean the "thermodynamic limit" $p \to \infty$. We consider the average error, the symbol $\mathbb{E}$ meaning the expectation with respect to the signal $\mathbf{x}$ and the data $\mathbf{y}$ (conditional on $\mathbf{x}$), seen as r.vs. drawn from their respective distributions (prior and likelihood). This quantity summarizes in a single number all the complexity of the high-d problem, by providing the optimal error one can aim for any algorithm, in average over all possible realisations of the problem.

One may wonder whether the number MMSE is sufficient to describe the problem, as it might a-priori be the case that the r.v. MMSE_p (random through $(\mathbf{x}, \mathbf{y})$) fluctuates a lot (i.e., has $O(1)$ variance). But in well-defined high-d inference problems in the Bayesian optimal setting it **concentrates**; it is said to be **self-averaging** in physics terminology. This means that it actually does not fluctuate much for large systems, and becomes deterministic in the limit $p, n \to \infty$:

$$\text{MMSE}_p = \text{MMSE} + o_p(1)$$

where $o_p(1)$ is by definition a quantity such that $\lim_p o_p(1) = 0$. Therefore it is sufficient to focus on the asymptotic averaged MMSE in order to capture/predict the behavior of fixed large (typical) instances of the problem. The self-averaging of error metrics in high-d Bayes-optimal inference is very generic [4, 5] (in mismatched settings this is not necessary the case). It is a non-trivial manifestation of the phenomenon of **concentration of measure** in high-d models, which is at the core of the determinism/predictability of these complex random systems.

High-dimensional inference as statistical mechanics

Let us establish now clear connections between what we discussed until now on high-d inference and statistical mechanics.

La posterior come distribuzione di Gibbs-Boltzmann

Il problema che abbiamo già accennato inerentemente il normalizzare una distribuzione di probabilità ad alta-d come quella a posteriori (cioè, il calcolo dell'evidenza) dovrebbe far suonare un campanello nei i fisici: la stessa cosa accade nella meccanica statistica, dove uno dei compiti principali è calcolare la **funzione di partizione** $\mathcal{Z}(\mathbf{J}) := \sum_{\boldsymbol{\sigma}} \exp\{-\beta \mathcal{H}(\boldsymbol{\sigma}; \mathbf{J})\}$ che normalizza la distribuzione di **Gibbs-Boltzmann**:

$$\mathbb{P}_{\text{GB}}(\boldsymbol{\sigma}; \mathbf{J}) = \tfrac{1}{\mathcal{Z}(\mathbf{J})} \exp\{-\beta \mathcal{H}(\boldsymbol{\sigma}; \mathbf{J})\}. \qquad (4)$$

Qui $\mathcal{H}(\boldsymbol{\sigma}; \mathbf{J})$ è la **Hamiltoniana / energia** che definisce il modello, e $\boldsymbol{\sigma}$ (spesso binario) sono gli **spin**. $\mathbf{J}$ sono i loro accoppiamenti **casuali e congelati**, cioè un insieme di variabili fisse che parametrizzano l'Hamiltoniana. β è la **temperatura inversa**.

Possiamo spingere ulteriormente l'analogia: la distribuzione a posteriori data dalla formula di Bayes (1) può essere naturalmente pensata come una distribuzione di Gibbs-Boltzmann nel contesto dell'inferenza ad alta-d. È sufficiente riscriverla in forma esponenziale e identificare le variabili (eventualmente a valori reali) $\mathbf{x}$ che rappresentano il segnale sconosciuto con gli spin $\boldsymbol{\sigma}$, e i dati $\mathbf{y}$ con le variabili casuali congelate $\mathbf{J}$:

$$\mathbb{P}(\mathbf{x} \mid \mathbf{y}) = \tfrac{1}{\mathcal{Z}(\mathbf{y})} \exp\{\ln \mathbb{P}(\mathbf{x}) + \ln \mathbb{P}(\mathbf{y} \mid \mathbf{x})\}. \quad (5)$$

Quindi la funzione di partizione è l'evidenza, e l'hamiltoniana è (meno) il logaritmo della prior più la log-verosimiglianza, mentre $\beta = 1$. In molti modelli interessanti la prior fattorizza sugli ingressi del segnale $\mathbb{P}(\mathbf{x}) = \prod_{i=1}^{p} P(x_i)$ (questi, cioè, sono i.i.d.), e la verosimiglianza anche fattorizza nei dati $\mathbb{P}(\mathbf{y} \mid \mathbf{x}) = \prod_{j=1}^{n} Q(y_j \mid \mathbf{x})$ (cioè, i dati sono condizionatamente i.i.d.). In questo caso recuperiamo una forma familiare per la posterior $\mathbb{P}(\mathbf{x} \mid \mathbf{y})$:

$$\tfrac{1}{\mathcal{Z}(\mathbf{y})} \exp\{\textstyle\sum_{i=1}^{p} \ln P(x_i) + \sum_{j=1}^{n} \ln Q(y_j \mid \mathbf{x})\}.$$

Quindi i termini locali $(\ln P(x_i))_{i=1}^{p}$ agiscono come campi magnetici esterni, mentre i termini di verosimiglianza $(\ln Q(y_j \mid \mathbf{x}))_{j=1}^{n}$ fungono da interazioni tra spin che li correlano in modo assolutamente non banale. Se fossero a coppie, re-

The posterior as a Gibbs-Boltzmann distribution

The problem that we already mentionned of normalizing a high-d probability distribution like the posterior (i.e., computing the evidence) should ring a bell for physicists: the very same thing happens in statistical mechanics, where one of the main task is to compute the **partition function** $\mathcal{Z}(\mathbf{J}) := \sum_{\boldsymbol{\sigma}} \exp\{-\beta \mathcal{H}(\boldsymbol{\sigma}; \mathbf{J})\}$ which normalizes the **Gibbs-Boltzmann distribution**:

$$\mathbb{P}_{\text{GB}}(\boldsymbol{\sigma}; \mathbf{J}) = \tfrac{1}{\mathcal{Z}(\mathbf{J})} \exp\{-\beta \mathcal{H}(\boldsymbol{\sigma}; \mathbf{J})\}. \qquad (4)$$

Here $\mathcal{H}(\boldsymbol{\sigma}; \mathbf{J})$ is the **Hamiltonian/energy function** defining the model, and $\boldsymbol{\sigma}$ (often binary) are the **spins**. $\mathbf{J}$ is the **quenched randomness**, namely, a set of fixed variables that parametrise the Hamiltonian. β is the **inverse temperature**.

We can push the analogy further: the posterior distribution given by the Bayes formula (1) can be naturally thought as a Gibbs-Boltzmann distribution in the context of high-d inference. Simply re-write it in exponential form and identify the (possibly real-valued) variables $\mathbf{x}$ representing the unknown signal with the spins $\boldsymbol{\sigma}$, and the data $\mathbf{y}$ with the quenched randomness $\mathbf{J}$:

$$\mathbb{P}(\mathbf{x} \mid \mathbf{y}) = \tfrac{1}{\mathcal{Z}(\mathbf{y})} \exp\{\ln \mathbb{P}(\mathbf{x}) + \ln \mathbb{P}(\mathbf{y} \mid \mathbf{x})\}. \quad (5)$$

So the partition function is the evidence, and the Hamiltonian is (minus) the log-prior plus log-likelihood, while $\beta = 1$. In many interesting models the prior factorises over the signal entries $\mathbb{P}(\mathbf{x}) = \prod_{i=1}^{p} P(x_i)$ (i.e., they are i.i.d.), and the likelihood factorises too over the data points $\mathbb{P}(\mathbf{y} \mid \mathbf{x}) = \prod_{j=1}^{n} Q(y_j \mid \mathbf{x})$ (i.e., the data points are conditionally i.i.d.). In this case we recover a familiar form for the posterior $\mathbb{P}(\mathbf{x} \mid \mathbf{y})$:

$$\tfrac{1}{\mathcal{Z}(\mathbf{y})} \exp\{\textstyle\sum_{i=1}^{p} \ln P(x_i) + \sum_{j=1}^{n} \ln Q(y_j \mid \mathbf{x})\}.$$

So the local terms $(\ln P(x_i))_{i=1}^{p}$ act as external magnetic fields, while the log-likelihood terms $(\ln Q(y_j \mid \mathbf{x}))_{j=1}^{n}$ are interactions between spins that correlate them in a highly non-trivial way. If these were pairwise, we would recover an instance of the **Ising model**, see below.

cupereremmo un'istanza del **modello di Ising**, (si veda a seguire).

Tornando alla nozione di stimatori ottimali: tenendo presente la nostra interpretazione della meccanica statistica, ora comprendiamo che la stima MAP è equivalente a **trovare lo stato fondamentale**, cioè la configurazione di spin che minimizza l'energia. Invece il calcolo dello stimatore MMSE si basa sul **campionamento della distribuzione a posteriori, o di Gibbs-Boltzmann**; questi sono tra i principali compiti algoritmici in meccanica statistica e corrispondono esattamente allo scopo dell'inferenza in oggetto.

All'improvviso diventa quasi ovvio che l'inferenza ad alta-d e la meccanica statistica siano cugini molto vicini: discuteremo esempi concreti. Tuttavia questa reinterpretazione non è semplicemente un'osservazione: consente di importare nel mondo dell'inferenza e nella data-science l'enorme quantità di tecniche e concetti di analisi sviluppati per più di un secolo nella meccanica statistica.

Modelli paradigmatici in meccanica statistica, e parametri d'ordine

Un riflesso della fisica che si è diffuso praticamente in tutte le aree scientifiche è quello di considerare un modello giocattolo contenente tutte le caratteristiche salienti di modelli più complessi/realistici, ma che sia al contempo abbastanza semplice da poter essere affrontato analiticamente al fine di comprendere i fenomeni fondamentali di più ampio respiro. Cominciamo con l'introduzione di due di questi modelli nella fisica statistica. Successivamente li collegheremo all'inferenza.

Il modello di gran lunga più comprensibile e più paradigmatico in meccanica statistica è il **modello di Ising completamente connesso**, chiamato anche **modello di Curie-Weiss (CW)**, definito dall'Hamiltoniana ($J > 0$, $h \in \mathbb{R}$ e $\boldsymbol{\sigma} \in \{-1,1\}^p$)

$$\mathcal{H}_{\text{CW}}(\boldsymbol{\sigma}; J, h) = -\frac{J}{p}\sum_{i<j}^{p}\sigma_i\sigma_j - h\sum_{i}^{p}\sigma_i.$$

La meccanica statistica usa quantità macroscopiche chiamate **parametri d'ordine** per descrivere un sistema complesso. Per il modello CW questo è semplicemente la **magnetizzazione** $m_p(\boldsymbol{\sigma}) := \frac{1}{p}\sum_{i=1}^{p}\sigma_i$. Quindi $m := \lim_p \langle m_p(\boldsymbol{\sigma})\rangle$ discerne se

Coming back to the notion of optimal estimators: with our statistical mechanics interpretation in mind, we now understand that MAP estimation is equivalent to **finding the ground state**, namely the spin configuration that minimizes the energy. Instead computing the MMSE estimator relies on **sampling the posterior/Gibbs-Boltzmann distribution**; these are among the main algorithmic tasks in statistical mechanics, and correspond exactly to the inference task.

Suddenly it becomes almost obvious that high-d inference and statistical mechanics are very close cousins. We will discuss concrete examples. Yet this re-interpretation is not simply an observation: it allows to import the massive amount of analysis techniques and concepts developed for more than a century in statistical mechanics into the world of inference and data-sciences.

Paradigmatic models of statistical mechanics, and order parameters

One physicist reflex that spreaded in virtually all scientific areas is to consider a toy model containing all the relevant features of more complex/realistic models, but yet being simple enough to be tackled analytically in order to understand fundamental phenomena that should apply more broadly. Let us start by introducing two such models in statistical physics. We will later connect them to inference.

By far the better understood and most paradigmatic model in statistical mechanics is the **fully-connected Ising model**, also called **Curie-Weiss model (CW)**, defined by the Hamiltonian ($J > 0$, $h \in \mathbb{R}$ and $\boldsymbol{\sigma} \in \{-1,1\}^p$)

$$\mathcal{H}_{\text{CW}}(\boldsymbol{\sigma}; J, h) = -\frac{J}{p}\sum_{i<j}^{p}\sigma_i\sigma_j - h\sum_{i}^{p}\sigma_i.$$

Statistical mechanics uses macroscopic quantities called **order parameters** for describing a complex system. For the CW model it is simply the **magnetisation** $m_p(\boldsymbol{\sigma}) := \frac{1}{p}\sum_{i=1}^{p}\sigma_i$. Then $m := \lim_p \langle m_p(\boldsymbol{\sigma})\rangle$ describes whether the system is in an ordered **ferromagnetic phase** (if $m \neq 0$) or a

il sistema sia in una **fase ferromagnetica** ordinata (se $m \neq 0$) o in una **fase ergodica** disordinata (se $m = 0$)[3]; qui abbiamo introdotto la notazione fisica standard $\langle \cdot \rangle$ per denotare una media rispetto alla distribuzione di Gibbs-Boltzmann (4). In questo semplice modello la concentrazione in misura implica $m_p(\boldsymbol{\sigma}) = m + o_p(1)$.

Un altro modello importante nei sistemi di spin è la versione disordinata del modello CW. Il **modello di Sherrington-Kirkpatrick (SK)**, o **vetro di spin in campo medio**, è definito dall'Hamiltoniana

$$\mathcal{H}_{\mathrm{SK}}(\boldsymbol{\sigma}; \mathbf{J}, h) = -\sum_{i<j}^{p} \frac{J_{ij}}{\sqrt{p}} \sigma_i \sigma_j - h \sum_{i}^{p} \sigma_i. \quad (6)$$

Le interazioni congelate $J_{ij} \sim \mathcal{N}(0,1)$ sono i.i.d. realizzazioni di una variabile casuale normale.

Apriamo una parentesi tecnica che non è cruciale per il resto della discussione. Un singolo parametro d'ordine scalare, quale la magnetizzazione, non è più sufficiente per descrivere la fenomenologia del modello SK. Invece è necessario considerare un parametro d'ordine distributivo più ricco $\mathbb{P}(q) := \lim_p \mathbb{E}_{\mathbf{J}} \mathbb{P}(q_p \mid \mathbf{J})$, che è la distribuzione di probabilità asintotica dell'**overlap** $q_p := \frac{1}{p} \sum_{i=1}^{p} \sigma_i^{(1)} \sigma_i^{(2)}$. Qui $\boldsymbol{\sigma}^{(1)}$ e $\boldsymbol{\sigma}^{(2)}$ sono (condizionatamente a $\mathbf{J}$) i.i.d. vettori casuali tratti dalla stessa distribuzione di Gibbs-Boltzmann: sono spesso chiamati repliche. Nel caso del modello CW la distribuzione asintotica della magnetizzazione era semplicemente una delta di Dirac δ_m, quindi m descriveva completamente il sistema. Ma qui la concentrazione in misura è molto più sottile: l'overlap non si concentra a bassa temperatura ($q_p \neq \lim_p \mathbb{E}_{\mathbf{J}} \langle q_p \rangle + o_p(1)$), ma la sua distribuzione: $\mathbb{P}(q_p \mid \mathbf{J})$ converge in distribuzione a $\mathbb{P}(q)$ come $p \to \infty$. La forma di questa distribuzione permette poi di descrivere le varie fasi del modello (ferromagnetica, ergodica, spin glass, ecc.). Pertanto la presenza del disordine $\mathbf{J}$ cambia drasticamente il modello e

[3] Questo è vero ogni volta che $h \neq 0$. In un sistema come il CW, a campo esterno nullo, dove è presente una simmetria globale per inversione del segno $\mathcal{H}_{\mathrm{CW}}(\boldsymbol{\sigma}; J, h = 0) = \mathcal{H}_{\mathrm{CW}}(-\boldsymbol{\sigma}; J, h = 0)$, è necessario rompere questa simmetria introducendo un piccolo campo esterno e quindi portando il limite di questo campo a 0 dopo aver eseguito il limite termodinamico. Il valore della magnetizzazione risultante $m^{\pm} := \lim_{h \to 0^{\pm}} \lim_p \frac{1}{p} \sum_{i=1}^{p} \langle \sigma_i \rangle_h$ dipenderà, al di sotto della sua temperatura critica $1/\beta_c$, a seconda che il limite $h \to 0$ sia preso dal basso o dall'alto.

disordered **antiferromagnetic phase** (if $m = 0$)[3]; here we introduced the standard physics notation $\langle \cdot \rangle$ to denote an average with respect to the Gibbs-Boltzmann distribution (4). In this simple model the concentration of measure implies $m_p(\boldsymbol{\sigma}) = m + o_p(1)$.

Another important model of spin system is the disordered version of the CW model. The **Sherrington-Kirkpatrick (SK) model**, or **mean-field spin glass**, is defined by the Hamiltonian

$$\mathcal{H}_{\mathrm{SK}}(\boldsymbol{\sigma}; \mathbf{J}, h) = -\sum_{i<j}^{p} \frac{J_{ij}}{\sqrt{p}} \sigma_i \sigma_j - h \sum_{i}^{p} \sigma_i. \quad (6)$$

The quenched interactions $J_{ij} \sim \mathcal{N}(0,1)$ are i.i.d. realisations of a normal random variable.

Let us open a technical parenthesis that is not crucial for the remaining of the discussion. A single scalar order parameter like the magnetisation is not enough anymore to describe the phenomenology of the SK model. Instead one needs to consider a richer distributional order parameter $\mathbb{P}(q) := \lim_p \mathbb{E}_{\mathbf{J}} \mathbb{P}(q_p \mid \mathbf{J})$, which is the asymptotic probability distribution of the **overlap** $q_p := \frac{1}{p} \sum_{i=1}^{p} \sigma_i^{(1)} \sigma_i^{(2)}$. Here $\boldsymbol{\sigma}^{(1)}$ and $\boldsymbol{\sigma}^{(2)}$ are (conditionally on $\mathbf{J}$) i.i.d. random vectors drawn from the same Gibbs-Boltzmann distribution; there are often called replicas. In the case of the CW model the asymptotic distribution of the magnetisation was simply a dirac mass δ_m, so m fully described the system. But here the concentration of measure is much more subtle: the overlap does not concentrate at low temperature ($q_p \neq \lim_p \mathbb{E}_{\mathbf{J}} \langle q_p \rangle + o_p(1)$), but its distribution does: $\mathbb{P}(q_p \mid \mathbf{J})$ converges in distribution to $\mathbb{P}(q)$ as $p \to \infty$. The shape of this distribution then allows to describe the various phases of the model (ferromagnetic, antiferromagnetic, spin glass, etc). Therefore the precense of disorder $\mathbf{J}$ changes drastically the model and its phenomenology, and describing it goes beyond the scope of this article, see [6, 7] for more de-

[3] This is true whenever $h \neq 0$. In a system with a global sign-flip symmetry like here when the external field is null $\mathcal{H}_{\mathrm{CW}}(\boldsymbol{\sigma}; J, h = 0) = \mathcal{H}_{\mathrm{CW}}(-\boldsymbol{\sigma}; J, h = 0)$, one needs to break this symmetry by introducing a small external field, and then taking the limit of this field to 0 after the thermodynamic limit. The resulting magnetisation value $m^{\pm} := \lim_{h \to 0^{\pm}} \lim_p \frac{1}{p} \sum_{i=1}^{p} \langle \sigma_i \rangle_h$ will depend, below its critical temperature $1/\beta_c$, on whether the limit $h \to 0$ is taken from below or above.

la sua fenomenologia, e descriverlo va oltre lo scopo di questo articolo, si veda [6, 7] per maggiori dettagli[4]. Il modello SK ha generato un intero campo di ricerca al crocevia tra fisica, teoria dell'informazione, informatica e matematica. E come ci accorgeremo presto, questo modello e il suo cugino non disordinato, il modello CW, sono entrambi profondamente connessi anche all'inferenza statistica.

Qual è il parametro d'ordine nell'inferenza ad alta-d? Un candidato naturale è una metrica di errore che ben si concentrerà sul MMSE (3). L'MMSE caratterizza le **fasi della teoria dell'informazione**: la fase di inferenza teoricamente possibile in cui l'MMSE è relativamente piccolo, ed il regime di inferenza impossibile dove è relativamente alto. Si noti che in generale la posizione della **transizione di fase nella teoria dell'informazione** che separa queste fasi non dipende da quale metrica di errore è usata per sondare il **diagramma di fase**; torneremo su queste nozioni.

Termodinamica: entropia libera e transizioni di fase

Una quantità chiave nella meccanica statistica è l'**entropia libera** (o "meno" l'**energia libera**):

$$f_p = \frac{1}{\beta p} \ln \mathcal{Z}.$$

Questa contiene tutte le informazioni termodinamiche sul modello: i punti di non analiticità del suo limite termodinamico $f := \lim_p f_p$ corrispondono alle posizioni delle **transizioni di fase**. Una transizione di fase avviene quando un sistema complesso sperimenta un cambiamento nel comportamento di certi parametri d'ordine al variare dei **parametri di controllo** esterni. L'esempio canonico è l'acqua, la cui fase può essere caratterizzata da una densità locale di molecole o da una lunghezza media di correlazione (due parametri d'ordine) durante il cambiamento della temperatura e/o della pressione (due parametri di controllo).

Uno dei vantaggi principali nell'uso dell'en-

[4] La soluzione del modello SK è stata trovata da G. Parisi [10, 12] e la dimostrazione rigorosa della soluzione proposta da Parisi ottenuta da F. Guerra [9] e M. Talagrand [8] (e successivamente riconfermata da D . Panchenko [7]).

tails[4]. The SK model has generated a whole field of research at the crossroad of physics, information theory, computer science and mathematics. And as we will realize soon, this model and its non-disordered cousin the CW model are both deeply connected to statistical inference too.

What is the order parameter in high-d inference? A natural candidate is an error metric, and well will focus on the MMSE (3). The MMSE characterizes the **information-theoretic phases**: information-theoretically possible inference phase where the MMSE is relatively small, and the impossible inference regime where it is comparatively high. Note that in general the location of the **information-theoretic phase transition** separating these phases does not depend on which error metric is used to probe the **phase diagram**; we will come back to these notions.

Thermodynamics: free entropy and phase transitions

A key quantity in statistical mechanics is the **free entropy** (or minus the **free energy**):

$$f_p = \frac{1}{\beta p} \ln \mathcal{Z}.$$

It contains all thermodynamic information about the model: the non-analyticity points of its thermodynamic limit $f := \lim_p f_p$ correspond to the location of **phase transitions**. A phase transition is when a complex system experiences a change in the behavior of certain order parameters when external **control parameters** are varied. The canonical example is water, whose phase may by characterized by a local density of molecules or an average correlation lenght (two order parameters) while the temperature and/or the pressure evolve (two control parameters).

One of the main use of the free entropy is that it allows to access (some) order parameters and

[4] The solution of the SK model has been found by G. Parisi [10, 12] and the rigorous proof of the Parisi solution obtained by F. Guerra [9] and M. Talagrand [8] (and later re-proved by D. Panchenko [7]).

tropia libera è che permette di accedere ad alcuni parametri d'ordine e alle loro fluttuazioni: questa è infatti la funzione generatrice dei momenti ed i parametri d'ordine sono proprio i momenti. Ad esempio, nel modello CW la magnetizzazione e le sue fluttuazioni sono ottenute prendendo le derivate rispetto al campo h:

$$f'_p = \langle m_p \rangle, \quad f''_p = p\langle (m_p - \langle m_p \rangle)^2 \rangle \qquad (7)$$

dove il simbolo $'$ significa derivata rispetto ad h. Per i sistemi disordinati, come il modello SK, generalmente consideriamo l'entropia libera attesa

$$\mathbb{E} f_p = \tfrac{1}{\beta p} \mathbb{E}_{\mathbf{J}} \ln \mathcal{Z}(\mathbf{J}).$$

Questa è equivalente all'entropia libera non mediata, ma è più pratica da calcolare poichè indipendente dalla particolare realizzazione delle interazioni $\mathbf{J}$. L'equivalenza è ancora una conseguenza della concentrazione in misura, che implica $f_p(\mathbf{J}) = \lim_p \mathbb{E} f_p + o_p(1)$. Si noti che anche se l'overlap non auto-media, come nel modello SK ed altri vetri di spin a bassa temperatura (o in problemi di ottimizzazione combinatoria [6]), l'energia libera è sempre auto-mediante (per ogni modello ben definito).

Approfondiamo la nozione di transizione di fase. Esistono molti tipi di transizione di fase; a volte sono abbastanza liscie (queste sono del tipo **secondo ordine** perché corrispondono a una discontinuità di una derivata di secondo ordine dell'entropia libera nel limite termodinamico), e talvolta molto nitide e discontinue (del **primo ordine**, cioè con una discontinuità di una derivata del primo ordine di f). Esempi di transizioni di fase sono: il recupero (retrieval) di un pattern da parte del cervello una volta che sono stati forniti sufficienti stimoli nella direzione del pattern, assunto memorizzato (un semplice modello di memoria associativa è il **modello di Hopfield**): qui il parametro d'ordine è la sovrapposizione della rete con il pattern memorizzato e il parametro di controllo è la quantità di stimoli. Una crepa nel mercato finanziario, dove improvvisamente tutti i prezzi scendono all'unisono. L'improvvisa transizione che si verifica quando si impacchettano casualmente abbastanza palline in una scatola (questa è chiamata **jamming transition**, ed è correlata all'ottimizzazione della memoria del

their fluctuations; it is the moment generating function, the order parameter(s) being the moments. E.g., in the CW model the magnetisation and its flucutuations are obtained by taking derivatives w.r.t. the field h:

$$f'_p = \langle m_p \rangle, \quad f''_p = p\langle (m_p - \langle m_p \rangle)^2 \rangle \qquad (7)$$

where the symbol $'$ means a h-derivative. For disordered systems like the SK model, we generally consider the expected free entropy

$$\mathbb{E} f_p = \tfrac{1}{\beta p} \mathbb{E}_{\mathbf{J}} \ln \mathcal{Z}(\mathbf{J}).$$

It is equivalent to the non-averaged free entropy, but more practical to compute as independent of a particular realisation of the interactions $\mathbf{J}$. The equivalence is again a consequence of the concentration of measure, that implies $f_p(\mathbf{J}) = \lim_p \mathbb{E} f_p + o_p(1)$. Note that even if the overlap is not self-averaging, like in the SK model and other spin glasses at low temperature (or combinatorial optimisation problems [6]), the free energy is always self-averaging (for well defined models).

Let us discuss further the notion of phase transition. There exist many types of phase transition; sometimes they are quite smooth (these are of the **second order** type because they correspond to a discontinuity of a second-order derivative of the asymptotic free entropy), and sometimes very sharp and discontinuous (of the **first order** type, namely, a discontinuity of a first order derivative of f). Examples of phase transitions are: the recovery of a souvenir by the brain once enough stimuli in the direction of the memorized pattern are provided (a simple model of associative memory is the **Hopfield model**). Here the order parameter is the overlap with the memorized pattern and the control parameter is the amount of stimuli. A crack in the financial market, where suddenly all prices drop all together. The sudden rigidity transition that happens when you randomly pack enough balls in a box (this is called the **jamming transition**, and this is related to computer memory optimization or error correcting codes in communication). When communicating bits through a given noisy channel, there

computer o ai codici di correzione degli errori nella comunicazione). Quando si comunicano bits attraverso un dato canale rumoroso, esiste una velocità massima di trasmissione di queste informazioni; la comunicazione al di sopra di questa soglia è impossibile poiché le informazioni vengono perse a causa del rumore. Questo limite è chiamato **capacità di Shannon** [13, 14] ed in realtà altro non è che una transizione di fase. Il parametro di ordine è la qualità del recupero dei bit di informazione, il parametro di controllo è la velocità di comunicazione. Un ultimo: supponiamo di voler addestrare un algoritmo di classificazione che, quando viene fornito un ampio database di esempi di addestramento etichettati, è in grado di distinguere le immagini di cani e gatti. Esiste un numero minimo di esempi di addestramento al di sotto dei quali, indipendentemente dalla potenza del computer, l'algoritmo non sarà mai in grado di classificare correttamente le immagini; questa è una transizione di teoria dell'informazione. Il parametro order è la prestazione di classificazione dell'algoritmo, il parametro di controllo è la dimensione del training set, si veda, ad esempio, [24, 25].

Teoria dell'informazione: entropia di Shannon entropy e mutua informazione

La teoria dell'informazione, nel contesto dell'inferenza, si occupa principalmente della seguente domanda: quando i dati contengono informazioni sufficienti da poter essere utilizzati per dedurre qualcosa sul processo che li ha generati?

Per affrontare questa domanda e, a corollario, capire quale sia la nozione cugina di entropia libera nell'inferenza ad alta-d, dobbiamo prima ricordare cosa sia l'**entropia di Shannon**. La comprensione di questo oggetto è fondamentale e ha dato origine alla nascita della teoria dell'informazione stessa [13], quindi dedicheremo del tempo a discuterne in dettaglio. Come inteso da Shannon, essa è correlata alla vecchia nozione di entropia in termodinamica e meccanica statistica, come vedremo, da cui il nome. La sua definizione, mediante v.c. discrete è

$$H(\mathbf{x}) := \sum_{\mathbf{x}} \mathbb{P}(\mathbf{x}) \ln \frac{1}{\mathbb{P}(\mathbf{x})} = \mathbb{E}_{\mathbf{x}} \ln \frac{1}{\mathbb{P}(\mathbf{x})}.$$

Abusiamo leggermente della notazione e usiamo lo stesso simbolo $\mathbf{x}$ per una v.c. e l'estrazione di

exists a maximum rate of information transmission; communication above this sharp threshold is impossible as information gets lost due to noise. This limit is called the **Shannon capacity** [13, 14] and really is nothing but a phase transition. The order parameter is the quality of recovery of the information bits, the control parameter is the communication rate. A final one. Say you want to train a classification algorithm that, when given a large data-base of labeled training examples, is able to distinguish pictures of dogs and cats. There exists a minimum number of training examples below which, no matter the power of the computer, the algorithm will never be able to properly classify the images; this is an information-theoretic transition. The order parameter is the classification performance of the algorithm, the control parameter is the size of the training set, see, e.g., [24, 25].

Information theory: Shannon entropy and mutual information

Information theory in the context of inference is mainly concerned with the following question: when does data contains enough information so that it can be used to infer something about the process that generated it?

To adress this question and, connected to that, understand what is the cousin notion of free entropy in high-d inference, we first need to recall what is the **Shannon entropy**. The understanding of this object is fundamental and gave rise to the birth of information theory [13], so we will spend some time to discuss it in details. As understood by Shannon, it is related to the older notion of entropy in thermodynamics and statistical mechanics as we will see, thus the name. Its definition for a discrete r.v is

$$H(\mathbf{x}) := \sum_{\mathbf{x}} \mathbb{P}(\mathbf{x}) \ln \frac{1}{\mathbb{P}(\mathbf{x})} = \mathbb{E}_{\mathbf{x}} \ln \frac{1}{\mathbb{P}(\mathbf{x})}.$$

We slightly abuse notation and use the same symbol $\mathbf{x}$ for a r.v. and its outcome (while information theory usually denotes the r.v. in capital

un suo valore (mentre la teoria dell'informazione di solito denota la v.c. in maiuscolo $\mathbf{X}$ ed un'estrazione in minuscolo $\mathbf{x}$). Limiteremo la nostra discussione a v.c. discrete poiché l'interpretazione è più sottile nel caso continuo, ma molta dell'intuizione vi si generalizza.

L'entropia di Shannon è l'attesa $\mathbf{x}$ del **contenuto informativo** $h(\mathbf{x}) := -\ln \mathbb{P}(\mathbf{x})$, o **sorpresa**, nel risultato $\mathbf{x}$. Se questo risultato ha una bassa probabilità, l'osservazione è abbastanza sorprendente e porta molte informazioni inattese: $h(\mathbf{x})$ è alta. Se invece $\mathbb{P}(\mathbf{x})$ è prossimo ad 1 non è sorprendente osservare $\mathbf{x}$, quindi questo risultato porta poca informazione: $h(\mathbf{x})$ è bassa. Detto diversamente: se l'esito di una v.c. è molto probabile, non è una sorpresa (e generalmente poco interessante) quando accade, perché era previsto. Tuttavia, se è improbabile che si verifichi un risultato, questo è molto più informativo se viene osservato. Il termine di contenuto informativo deve essere inteso come un *potenziale* guadagno di informazioni se si osserva $\mathbf{x}$. Quando si utilizza $\log_2$, il contenuto delle informazioni e l'entropia sono espressi in " bits".

Si immagini di essere nel deserto e improvvisamente piova a dirotto. Peggio ancora, piovono mucche che suonano il piano! Che cosa?! È sorprendentemente sorprendente no? La probabilità di questo evento è in realtà così bassa da portare con se un'enorme quantità di informazioni: in questo caso si dovrebbe raggiungere la conclusione che si stia sognando. Se invece si è in un deserto oltremodo soleggiato e torrido, non ci si sorprende affatto: questo evento non porta con se più informazioni di quelle già note, e se si sta sognando, è improbabile che questa osservazione aiuti a comprenderlo. Un altro esempio: la consapevolezza che un determinato numero di una lotteria non sarà quello vincente fornisce pochissime informazioni, perché qualsiasi numero scelto in particolare quasi certamente non vincerà. Tuttavia, la conoscenza che un determinato numero vincerà una lotteria ha un alto valore informativo perché comunica il risultato di un evento a probabilità molto bassa.

L'entropia può anche essere interpretata come una *misura dell'imprevedibilità* della v.c. $\mathbf{x}$, o di *disinformazione/mancanza di conoscenza* su quale sarà il risultato di $\mathbf{x}$: più sorprendenti sono i risultati nell'aspettativa, più imprevedibile è

letter $\mathbf{X}$ and an outcome $\mathbf{x}$). We will restrict our discussion to discrete r.vs. as the interpretation is a bit more subtle in the continuous case, but a lot of the intuition generalizes.

The Shannon entropy is the $\mathbf{x}$-expectation of the **information content** $h(\mathbf{x}) := -\ln \mathbb{P}(\mathbf{x})$, or **surprise**, of the outcome $\mathbf{x}$. If this outcome has low probability then observing it is quite surprising, and it brings a lot of information as it was not expected: $h(\mathbf{x})$ is high. If instead $\mathbb{P}(\mathbf{x})$ is close to 1 it is not surprising to observe $\mathbf{x}$, so this outcome brings low information: $h(\mathbf{x})$ is low. Said differently: if the outcome of a r.v. is very probable, it is no surprise (and generally uninteresting) when it happens, because it was expected. However, if an outcome is unlikely to occur, it is much more informative if it happens to be observed. The term information content must be understood as a *potential* information gain if $\mathbf{x}$ is observed. When using the $\log_2$ the information content and entropy are expressed in "bits".

Imagine you are in the desert and suddenly it rains like hell. Worst, it rains cows that play piano! What?! It is amazingly surprising no? The probability of this event is actually so low that it brings an enormous amount of information; in this case it should lead you to the conclusion that you are dreaming. If instead your are in the desert and its super sunny and hot, it is not surprising at all; this does not bring more information than what you already know, and if you are dreaming, it is unlikely that this observation will help you realize it. Another example: the knowledge that some particular number will not be the winning one of a lottery provides very little information, because any particular chosen number will almost certainly not win. However, knowledge that a particular number will win a lottery has high informational value because it communicates the outcome of a very low probability event.

The entropy can also be interpreted as a *measure of unpredictability* of the r.v. $\mathbf{x}$, or of *uninformation/lack of knowledge* about what $\mathbf{x}$'s outcome will be: the more surprising are the outcomes in expectation, the more unpredictable is the ac-

il risultato effettivo, il che significa anche che sappiamo meno di $\mathbf{x}$ *prima* di osservarlo. $H(\mathbf{x})$ *quantifica la quantità attesa di informazioni mancanti necessarie per determinare il risultato di* $\mathbf{x}$ *prima di osservarlo*. Questo può creare confusione perché in precedenza abbiamo detto che $H(\mathbf{x})$ è un contenuto informativo atteso, mentre ora parliamo di una misura di disinformazione. Non vi è alcun paradosso: un'informazione $H(\mathbf{x})$ è *acquisita* in media quando il risultato di $\mathbf{x}$ *viene effettivamente osservato*. Ma *prima* di osservare il risultato, $H(\mathbf{x})$ è una misura della disinformazione al riguardo. In altre parole: l'osservare il risultato $\mathbf{x}$ *converte* in media $H(\mathbf{x})$ unità di disinformazione in informazione. Quindi è solo questione di posizionarci concettualmente *prima* che $\mathbf{x}$ venga osservato –nel qual caso l'interpretazione come misura della disinformazione può essere più naturale–, o *dopo* che $\mathbf{x}$ sia osservato –dove l'interpretazione come contenuto informativo atteso sembra adattarsi meglio: alla fine è la stessa cosa.

Un esempio potrebbe aiutare: il risultato del lancio di una moneta equa $x_{\text{onesta}} \sim \text{Ber}(1/2)$ è molto più imprevedibile del risultato di una moneta fortemente truccata $x_{\text{truccata}} \sim \text{Ber}(9/10)$, o equivalentemente la nostra mancanza di conoscenza su cosa sarà x_{onesta} è maggiore: siamo più disinformati. Ma quando *osserviamo* il risultato della moneta equa, allora *otteniamo* più informazioni che con quella truccata, perché in media è più sorprendente. Nel primo caso, che ha entropia $H(x_{\text{onesta}})$ di un bit, scommettere su un lato o sull'altro è statisticamente identico. Mentre nel secondo caso, dove $H(x_{\text{truccata}}) = \frac{9}{10}\log_2\frac{10}{9} + \frac{1}{10}\log_2 10 \approx 0.47$, il risultato è molto più prevedibile, siamo meno disinformati (= più informati); sarebbe un errore non scommettere sul risultato $x_{\text{bias}} = 1$.

Riassumendo: l'entropia di Shannon $H(\mathbf{x})$ della v.c. $\mathbf{x}$ quantifica: $i)$ il suo contenuto informativo medio, cioè il guadagno di informazione atteso quando si osserva il risultato $\mathbf{x}$; $ii)$ la media disinformazione/mancanza di conoscenza del risultato $\mathbf{x}$ prima di osservarlo; $iii)$ la sua imprevedibilità. Maggiore è l'entropia di $\mathbf{x}$, meno "strutturata" è la sua distribuzione; $v)$ quando espresso in bit $H(\mathbf{x})$ è il numero atteso di domande binarie "sì/no" necessarie per determinare il risultato *prima* che si osserva o, equivalentemente, il numero atteso di domande binarie a cui

tual outcome, which also mean the less we know about $\mathbf{x}$ *before* observing it. $H(\mathbf{x})$ *quantifies the expected amount of missing information necessary to determine the outcome of* $\mathbf{x}$ *before observing it*. This can be confusing because previously we said that $H(\mathbf{x})$ is an expected information content, while now we speak about a measure of uninformation. There is no paradox: an information $H(\mathbf{x})$ is *gained* in average when $\mathbf{x}$'s outcome *is actually observed*. But *prior* to observing the outcome, $H(\mathbf{x})$ is a measure of uninformation about it. Put differently: observing the outcome $\mathbf{x}$ *converts* in average $H(\mathbf{x})$ units of uninformation into information. So it just a matter of conceptually placing ourselves *before* $\mathbf{x}$ is observed –in which case the interpretation as a measure of uninformation may be more natural–, or *after* $\mathbf{x}$ is observed –where the interpretation as an expected information content seems to fit better. But at the end this is the same thing.

An example might help: the outcome of a toss of a fair coin $x_{\text{fair}} \sim \text{Ber}(1/2)$ is much more unpredictable than the outcome of a strongly biased coin $x_{\text{bias}} \sim \text{Ber}(9/10)$, or equivalently our lack of knowledge about what will be x_{fair} is higher: we are more uninformed. But when *observing* the outcome of the fair coin, we then *gain* more information than with the unfair one, because it is in average more surprising. In the first case, which has entropy $H(x_{\text{fair}})$ of one bit, betting on one side or the other is the same statistically. While in the second case, where $H(x_{\text{bias}}) = \frac{9}{10}\log_2\frac{10}{9} + \frac{1}{10}\log_2 10 \approx 0.47$, the outcome is much more predictable, we are less uninformed (= more informed); it would be an error not to bet on the outcome $x_{\text{truccata}} = 1$.

To summarize: the Shannon entropy $H(\mathbf{x})$ of the r.v. $\mathbf{x}$ quantifies: $i)$ its average information content, i.e., the expected information gain when observing outcome $\mathbf{x}$; $ii)$ the average uninformation/lack of knowledge about the outcome $\mathbf{x}$ prior to observe it; $iii)$ its unpredictability. The higher the entropy of $\mathbf{x}$, the less "structured" its distribution is; $v)$ when expressed in bits $H(\mathbf{x})$ is the expected number of binary "yes/no" questions required to determine the outcome *before* it is observed, or equivalently, the expected number of binary questions that the oucome $\mathbf{x}$ has

l'evento $\mathbf{x}$ ha risposto *dopo* che è stato osservato.

Allo stesso modo l'**entropia condizionata** è:

$$H(\mathbf{x} \mid \mathbf{y}) := \sum_{\mathbf{x},\mathbf{y}} \mathbb{P}(\mathbf{y})\mathbb{P}(\mathbf{x} \mid \mathbf{y}) \ln \tfrac{1}{\mathbb{P}(\mathbf{x}|\mathbf{y})},$$

e quantifica le informazioni attese rivelate valutando il risultato di $\mathbf{x}$ assunto che si conosca già il risultato di $\mathbf{y}$, O, in modo equivalente, è la quantità attesa rimanente di imprevedibilità di $\mathbf{x}$ dato che $\mathbf{y}$ è già stato osservato.

L'entropia ha molte proprietà importanti che la rendono una "buona" definizione di contenuto informativo, una delle principali è che è additiva per v.c. indipendenti: $H(\mathbf{x}, \mathbf{y}) = H(\mathbf{y}) + H(\mathbf{x})$ se $\mathbb{P}(\mathbf{x}, \mathbf{y}) = \mathbb{P}(\mathbf{x})\mathbb{P}(\mathbf{y})$ e molte altre come la sua non-negatività (per v.c. discrete) e la regola della catena $H(\mathbf{x}, \mathbf{y}) = H(\mathbf{x} \mid \mathbf{y}) + H(\mathbf{y}) = H(\mathbf{y} \mid \mathbf{x}) + H(\mathbf{x})$. Ma tutte queste giustificazioni non sono sufficienti per *provare* che sia effettivamente *la* definizione corretta; forse altre funzioni verificano tutte queste proprietà e hanno un'interpretazione simile. La prova matematica che l'entropia è davvero la definizione corretta viene dal **teorema della codifica di sorgente** di C. Shannon, il padre della teoria dell'informazione, si veda [13, 2, 14]. Descriviamolo a parole, consideriamo i simboli binari ma il seguente ragionamento si applica agli alfabeti discreti più generici:

Approssimativamente, il teorema della codifica di sorgente afferma che, se una sorgente genera stringhe $(x_1, x_2, \ldots, x_n)$ di $n \gg 1$ simboli binari che sono i.i.d. risultati di una variabile casuale x, allora esiste un **codice** $\mathcal{C}_\delta$ compresso per questa sorgente di cardinalità $|\mathcal{C}_\delta| \approx 2^{nH(x)} \leq 2^n$, e questo indipendentemente dal rischio $0 < \delta < 1$ di perdere informazioni durante la codifica (nel tendere di $n \to \infty$).

Cerchiamo di capire cosa significhi e perché implichi che $H(x)$ sia la definizione corretta del contenuto informativo portato dalla v.c. x. $i)$ Primo, perché introdurre una sorgente di stringhe lunghe? Il contenuto informativo della sorgente, *qualunque cosa significhi*, deve essere n moltiplicato solo per x perché le informazioni devono essere additive per variabili indipendenti $(x_1, x_2, \ldots, x_n)$. Di conseguenza, il contenuto informativo atteso per simbolo della sorgente è uguale a quello di x. Quindi studiare la sorgente o studiare x è lo stesso da un punto di vista di teoria dell'informazione. Shannon comprese che,

answered *after* being observed.

Similarly the **conditional entropy** is:

$$H(\mathbf{x} \mid \mathbf{y}) := \sum_{\mathbf{x},\mathbf{y}} \mathbb{P}(\mathbf{y})\mathbb{P}(\mathbf{x} \mid \mathbf{y}) \ln \tfrac{1}{\mathbb{P}(\mathbf{x}|\mathbf{y})}.$$

It is the expected information revealed by evaluating the outcome of $\mathbf{x}$ given that you know already the outcome of $\mathbf{y}$. Or equivalently, it is the expected remaining amount of unpredictability of $\mathbf{x}$ given that $\mathbf{y}$ has already been observed.

The entropy has many important properties that make it a "good" definition of information content, one of the main being that it is additive for independent r.vs.: $H(\mathbf{x}, \mathbf{y}) = H(\mathbf{y}) + H(\mathbf{x})$ if $\mathbb{P}(\mathbf{x}, \mathbf{y}) = \mathbb{P}(\mathbf{x})\mathbb{P}(\mathbf{y})$, and many other ones such as its non-negativity (for discrete r.vs.) and the chain rule $H(\mathbf{x}, \mathbf{y}) = H(\mathbf{x} \mid \mathbf{y}) + H(\mathbf{y}) = H(\mathbf{y} \mid \mathbf{x}) + H(\mathbf{x})$. But all these justifications are not enough to *prove* that it is indeed *the* correct definition. Maybe other functions verify all these properties and have a similar interpretation. The mathematical proof that the entropy indeed is the correct definition comes from the **source coding theorem** of C. Shannon, the father of information theory, see [13, 2, 14]. Let us describe it in words. We consider binary symbols but the following reasonning applies to more generic discrete alphabets.

Roughly, the source coding theorem says that if a source generates strings $(x_1, x_2, \ldots, x_n)$ of $n \gg 1$ binary symbols that are i.i.d. outcomes of some random variable x, then there exists a compressed **code** $\mathcal{C}_\delta$ for this source of cardinal $|\mathcal{C}_\delta| \approx 2^{nH(x)} \leq 2^n$, and this independently of the risk $0 < \delta < 1$ we are ready to take in losing information when coding (as $n \to \infty$).

Let us understand what does that mean, and why it implies that $H(x)$ is the proper definition of information content carried by the r.v. x. $i)$ First, why introducing a source of long strings? The information content of the source, *whatever it means*, must be n times the one of x alone because information must be additive for independent variables $(x_1, x_2, \ldots, x_n)$. As a consequence the expected information content per symbol of the source equals the one of x. So studying the source or x is the same from an information-theoretic point of view. But as n will get large, Shannon understood that the concentration of measure in the form of the law of large num-

al crescere di n, la concentrazione in misura nella forma della legge dei grandi numeri avrebbe aiutato molto nell'analisi. $ii)$ Cos'è il codice? Un codice $\mathcal{C}_\delta$ è una qualsiasi rappresentazione alternativa "massimamente compressa" dell'insieme di stringhe in esame. Vale a dire che è un insieme di cardinalità inferiore a 2^n, il numero di stringhe possibili, in modo tale che una stringa casuale della sorgente abbia un elemento associato nel codice con probabilità (sulle stringhe) almeno $1-\delta$. E allo stesso tempo il codice abbi la cardinalità più piccola possibile. Un codice $\mathcal{C}_\delta$ quindi "codifica" parte delle informazioni della sorgente (sintonizzatocisi mediante δ) attraverso una relazione biiettiva tra $\mathcal{C}_\delta$ e un sottoinsieme delle possibili stringhe 2^n. Il rischio che corriamo è nel senso che con probabilità $< \delta$ una stringa generata dalla sorgente non avrà un elemento associato in $\mathcal{C}_\delta$, quindi le sue informazioni andranno perse durante la codifica. Un modo costruttivo per definire $\mathcal{C}_\delta$ è classificare tutte le possibili stringhe in base alla loro probabilità. Si aggiunge un primo elemento C_1 in $\mathcal{C}_\delta$, associato alla stringa più probabile, quindi si aggiunge un secondo elemento C_2 in $\mathcal{C}_\delta$ legato alla seconda stringa più probabile, e così via, fino a quando la somma delle probabilità delle stringhe relazionate agli elementi del codice supera $1-\delta$. $iii)$ Il contenuto informativo di questo codice espresso in bit è naturalmente definito come il numero di simboli binari necessari per rappresentare qualsiasi elemento del codice: $\log_2 |\mathcal{C}_\delta|$. Inoltre Shannon ha mostrato che $\log_2 |\mathcal{C}_\delta| \to nH(x)$ come $n \to \infty$. *A-priori* questo contenuto informativo è inferiore a quello della sorgente originale in quanto si è corso qualche rischio δ durante la compressione/mappatura della sorgente in $\mathcal{C}_\delta$; parliamo di **compressione con perdita**. $iv)$ Ma l'osservazione cruciale è che la quantità $H(x)$ che definisce la cardinalità del codice necessario per comprimere la sorgente fino al rischio $0 < \delta < 1$ diventa *indipendente* dal rischio nel limite $n \gg 1$. Ciò significa che fino a quando ci concediamo una piccola probabilità di errore δ (indipendente da n), è possibile una compressione fino a $nH(x)$ bits. Ma anche se ci è consentita una grande probabilità di errore, possiamo comunque comprimere la sorgente solo fino a $nH(x)$ bits: ciò suggerisce fortemente che $nH(x)$ è il contenuto informativo fondamentale della sorgente. Come conseguenza di questo

bers will help a lot in the analysis. $ii)$ What is a code? A code $\mathcal{C}_\delta$ is any alternative "maximally compressed" represention of the set of strings. Namely it is a set of smaller cardinal than 2^n, the number of possible strings, such that a random string from the source has an associated element in the code with probability (over the strings) at least $1-\delta$. And at the same time the code has smallest possibe cardinal. A code $\mathcal{C}_\delta$ therefore "encodes" part of the information (as tuned by δ) about the source through a bijective mapping between $\mathcal{C}_\delta$ and a subset of the 2^n possible strings. The risk we take is in the sense that with probability $< \delta$ a string generated by the source will not have an associated element in $\mathcal{C}_\delta$ so its information is lost when coding. A constructive way of defining $\mathcal{C}_\delta$ is to rank all possible strings according to their probability. Then add a first element C_1 in $\mathcal{C}_\delta$, associated to the most probable string. Then add a second element C_2 in $\mathcal{C}_\delta$ mapped to the second most probable string, and so on, until the sum of probabilities of the strings mapped to the code elements exceeds $1-\delta$. $iii)$ The information content of this code expressed in bits is naturally defined as the number of binary symbols necessary to represent any element of the code: $\log_2 |\mathcal{C}_\delta|$. Moreover Shannon showed that $\log_2 |\mathcal{C}_\delta| \to nH(x)$ as $n \to \infty$. *A-priori* this information content is less than the one of the original source as some risk δ has been taken when compressing/mapping the source to $\mathcal{C}_\delta$; we talk about **lossy compression**. $iv)$ But the crucial observation is that the quantity $H(x)$ defining the cardinal of the code necessary to compress the source up to risk $0 < \delta < 1$ becomes *independent* of the risk in the limit $n \gg 1$. This means that as long as we allow ourselves a tiny probability of error δ (independent of n), compression down to $nH(x)$ bits is possible. But even if we are allowed a large probability of error, we still can compress the source only down to $nH(x)$ bits. This strongly suggests that $nH(x)$ is the fundamental information content of the source. As a consequence of this and point $i)$ the information content of x is also $\frac{1}{n}\log_2 |\mathcal{C}_\delta| \to H(x)$. This ends the reasoning.

e del punto i), il contenuto informativo di x è anche $\frac{1}{n}\log_2|\mathcal{C}_\delta| \to H(x)$. Questo pone fine al ragionamento.

Diamo una dimostrazione ad alto livello del teorema di compressione di Shannon. Il punto è che, man mano che n diventa sempre più grande, per la legge dei grandi numeri quasi tutte le stringhe effettivamente generate dalla sorgente casuale sono *tipiche*, così che solo le sequenze tipiche devono essere codificate durante la compressione (le altre sono troppo improbabili e quindi non vengono codificate). Consideriamo per semplicità le variabili di Bernoulli $x \sim \text{Ber}(\rho)$: tutte le sequenze tipiche hanno approssimativamente lo stesso numero $n\rho$ di uno e $n(1-\rho)$ di zeri. Infatti, la probabilità che la stringa abbia esattamente R uno è una distribuzione binomiale $R \sim \text{Bin}(n,\rho)$. La fluttuazione relativa di R è $O(1/\sqrt{n})$ quindi R si concentra sulla sua media quando n diventa grande[5]. Ciò implica che, con alta probabilità, le uniche stringhe osservabili sono quelle con valori di R molto vicini a $n\rho$: questo definisce in modo informale l'*insieme tipico*. Quindi la probabilità di una sequenza tipica $\mathbf{x}_{\text{tipico}} = (x_1,\dots,x_n)$ è

$$\mathbb{P}(\mathbf{x}_{\text{tipico}}) = \textstyle\prod_i^n P(x_{\text{tipico},i}) \approx \rho^{n\rho}(1-\rho)^{n(1-\rho)}.$$

Chiamiamo questa probabilità di una stringa tipica $P_{\text{tipico}} := \rho^{n\rho}(1-\rho)^{n(1-\rho)}$. Qual è il contenuto/sorpresa dell'informazione, in bits, di un risultato tipico?

$$\begin{aligned}\log_2 \tfrac{1}{P_{\text{tipico}}} &= -n(\rho\log_2\rho + (1-\rho)\log_2(1-\rho)) \\ &= nH(x). \end{aligned} \tag{8}$$

Quindi la strategia dimostrativa è: i) man mano che n diventa grande si osservano solo sequenze/risultati tipiche/i; questi convogliano quasi tutta la massa di probabilità. Quindi, quando si definisce il codice $\mathcal{C}_\delta$, dobbiamo solo codificare questi risultati tipici; così facendo avviene la massima compressione. Il numero di stringhe tipiche è esponenzialmente grande in

[5]Fluttuazioni relative dell'ordine di $O(1/\sqrt{n})$ di grandezze macroscopiche come R sono tipiche dei sistemi complessi trattati in meccanica statistica. Il fatto che le fluttuazioni relative svaniscano è il motivo per cui tali sistemi casuali possono essere analizzati e descritti asintoticamente (quando $n \to +\infty$) da osservabili deterministiche, convergenti sulla loro media d'ensemble.

Let us give a high-level proof of the source coding theorem. The point is that, as n gets larger, by the law of large numbers almost all strings actually generated by the random source are *typical*, so that only the typical sequences need to be encoded during the compression (the others are too unprobable and therefore are not coded). Let us consider for simplicity Bernoulli variables $x \sim \text{Ber}(\rho)$. All typical sequences have approximately the same number $n\rho$ of ones and $n(1-\rho)$ of zeros. Indeed, the probability that the string has exactly R ones is a binomial distribution $R \sim \text{Bin}(n,\rho)$. The relative fluctuation of R is $O(1/\sqrt{n})$ so R concentrates onto its mean when n gets large[5]. This implies that with high probability the only possibly observed strings are those with R values very close to $n\rho$: this informally defines the *typical set*. So the probability of a typical sequence $\mathbf{x}_{\text{typ}} = (x_1,\dots,x_n)$ is

$$\mathbb{P}(\mathbf{x}_{\text{typ}}) = \textstyle\prod_i^n P(x_{\text{typ},i}) \approx \rho^{n\rho}(1-\rho)^{n(1-\rho)}.$$

Denote this probability of a typical string $P_{\text{typ}} := \rho^{n\rho}(1-\rho)^{n(1-\rho)}$. What is the information content/surprise in bits of a typical outcome?

$$\begin{aligned}\log_2 \tfrac{1}{P_{\text{typ}}} &= -n(\rho\log_2\rho + (1-\rho)\log_2(1-\rho)) \\ &= nH(x). \end{aligned} \tag{8}$$

So the proof strategy is: i) as n gets large only typical sequences/outcomes are observed; they carry almost all the probability mass. So when defining the code $\mathcal{C}_\delta$ we need only to code these typical outcomes; doing so it is maximally compressed. The number of typical strings is exponentially large in n (this follows from the **asymptotic equipartition principle**), so even if we allow a risk δ very close to 1 (but independent of n) and therefore only code a small fraction of the typical sequences, there are still approximately as many at leading (exponential) order

[5]Relative fluctuations of the order $O(1/\sqrt{n})$ of macroscopic quantities like R are typical of complex systems treated in statistical mechanics. That the relative fluctuations vanish is the reason why such random systems can be analyzed and described by asymptotically (as $n \to +\infty$) deterministic observables, converging on their ensemble mean.

n (questo segue dal **principio di equipartizione asintotica**), quindi anche se ci permettiamo un rischio δ molto vicino ad 1 (ma indipendente da n) e quindi codifichiamo solo una piccola frazione delle sequenze tipiche, ce ne sono ancora approssimativamente altrettante (espandendo ai termini dominanti esponenziali) quanto n diventa grande. Ad esempio, se ci sono $\exp(an)$ sequenze tipiche e codifichiamo solo $(1-\delta)\exp(an) = \exp(an+\ln(1-\delta))$, c'è lo stesso numero di sequenze al primo ordine d'espansione per qualsiasi $a > 0$ e $1 > \delta > 0$ fisso al crescere di $n \gg 1$. Quindi, indipendentemente da δ, il numero $|\mathcal{C}_\delta|$ di sequenze tipiche necessarie per la codifica è lo stesso, al primo ordine d'espanzione esponenziale. $ii)$ La domanda quindi diventa: possiamo contarle, cioè valutare $|\mathcal{C}_\delta|$ al primo ordine? Per definizione tutte le sequenze tipiche hanno approssimativamente la stessa probabilità P_{tipico} e trasportano quasi tutta la massa. Perciò

$$\sum_{\{\mathbf{x}\,\text{tipico}\}} P(\mathbf{x}) \approx \#_{\text{tipico}} P_{\text{tipico}} \approx 1,$$

dove $\#_{\text{tipico}}$ è il numero di sequenze tipiche. Per quanto detto in precedenza $\#_{\text{tipico}}$ è uguale a $|\mathcal{C}_\delta|$ al primo ordine. Ciò implica che ci sono approssimativamente $\#_{\text{tipico}} \approx 1/P_{\text{tipico}} = 2^{nH(X)}$ sequenze tipiche (da (8)) e possiamo quindi contarle espandendo al primo ordine. Ciò consente di stimare il contenuto informativo atteso per bit come $\frac{1}{n}\log_2|\mathcal{C}_\delta| \approx H(x)$, che è lo stesso del contenuto informativo atteso di x per la definizione della sorgente. Lo stesso argomento si estende ad alfabeti più generale (non binari).

Una quantità a questa connessa nella teoria dell'informazione è la **mutua informazione**:

$$I(\mathbf{x};\mathbf{y}) := H(\mathbf{x}) - H(\mathbf{x} \mid \mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y} \mid \mathbf{x}).$$

Questa viene interpretata come una misura della dipendenza reciproca di $\mathbf{x}$ e $\mathbf{y}$: quantifica la "quantità di informazione " ottenuta su una v.c. attraverso l'osservazione dell'altra. E infatti si annulla se e solo se le v.c. sono indipendenti: $I(\mathbf{x};\mathbf{y}) \geq 0$ con l'uguaglianza se e solo se $\mathbb{P}(\mathbf{x},\mathbf{y}) = \mathbb{P}(\mathbf{x})\mathbb{P}(\mathbf{y})$.

Per un problema di inferenza in cui vogliamo recuperare i parametri $\mathbf{x}$ dai dati $\mathbf{y}(\mathbf{x})$ l'ultima forma ha un'interpretazione particolarmente allettante: $H(\mathbf{y}) - H(\mathbf{y} \mid \mathbf{x})$ è l'informazione totale trasportata dai dati meno la rimanente im-

as n gets large. E.g., if there are $\exp(an)$ typical sequences and we only code $(1-\delta)\exp(an) = \exp(an + \ln(1-\delta))$ of them, there are the same number at leading order for any fixed $a > 0$ and $1 > \delta > 0$ as $n \gg 1$. So independently of δ the number $|\mathcal{C}_\delta|$ of typical sequences necessary to code is the same at leading exponential order. $ii)$ The question then becomes: can we count them, i.e., evaluate $|\mathcal{C}_\delta|$ at leading order? By definition all typical sequences have approximately the same probability P_{typ}, and they carry almost all the mass. Therefore

$$\sum_{\{\mathbf{x}\,\text{typical}\}} P(\mathbf{x}) \approx \#_{\text{typ}} P_{\text{typ}} \approx 1,$$

where $\#_{\text{typ}}$ is the number of typical sequences. With what we said previously $\#_{\text{typ}}$ equals $|\mathcal{C}_\delta|$ at leading order. This implies that there are approximately $\#_{\text{typ}} \approx 1/P_{\text{typ}} = 2^{nH(X)}$ typical sequences (from (8)). We can thus count them at leading order. This allows to estimate the expected information content per bit as $\frac{1}{n}\log_2|\mathcal{C}_\delta| \approx H(x)$, which is the same as the expected information content of x by definition of the source. The same argument extends to more general (non binary) alphabet.

A connected information-theoretic quantity is the **mutual information**:

$$I(\mathbf{x};\mathbf{y}) := H(\mathbf{x}) - H(\mathbf{x} \mid \mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y} \mid \mathbf{x}).$$

It is interpreted as a measure of the mutual dependence of $\mathbf{x}$ and $\mathbf{y}$. It quantifies the "amount of information" obtained about one r.v. through observing the other one. And indeed it cancels if and only if the r.vs. are independent: $I(\mathbf{x};\mathbf{y}) \geq 0$ with equality if and only if $\mathbb{P}(\mathbf{x},\mathbf{y}) = \mathbb{P}(\mathbf{x})\mathbb{P}(\mathbf{y})$.

In an inference problem where we want to recover the parameters $\mathbf{x}$ from the data $\mathbf{y}(\mathbf{x})$ the last form has a particularly nice interpretation: $H(\mathbf{y}) - H(\mathbf{y} \mid \mathbf{x})$ is the total information carried by the data minus the remaining unpredictability/uninformation about the data when the sig-

prevedibilità/disinformazione sui dati quando il segnale è noto, che è quindi il contributo del rumore. Ad esempio, in un modello gaussiano **denoising** $y = \sqrt{\lambda}\, x + z$ con $z \sim \mathcal{N}(0,1)$ abbiamo $H(y \mid x) = H(z) = \ln(2\pi e)/2$. L'informazione reciproca è quindi l'informazione trasportata dai dati di pura provenienza dal segnale. In quanto tale, quantifica i limiti dell'inferenza nella teoria dell'informazione e calcolarla in contesti ad alta-d è un obiettivo chiave della teoria dell'informazione stessa. Nel modello di denoising gaussiano è solo un esercizio mostrare che la sua espressione esplicita si legge (qui x^*, x sono i.i.d. da $\mathbb{P}(x)$)

$$I(x;y) = \tfrac{\lambda}{2}\mathbb{E}[x^2] - \mathbb{E}_{x^*} \ln \mathbb{E}_x e^{\lambda x^* x + \sqrt{\lambda} z x - \frac{\lambda}{2} x^2}. \quad (9)$$

Il denoising, la formula I-MMSE e l'interpretazione in teoria dell'informazione dell'entropia libera

Consideriamo il problema di denoising generale, dove $\mathbf{x}$ può essere un vettore, una matrice, ecc: $\mathbf{y} = \sqrt{\lambda}\,\mathbf{x} + \mathbf{z}$. La v.c. $\mathbf{z}$ ha la stessa dimensione del segnale ed ha ingressi normali standard i.i.d. Il RSR $\lambda > 0$ controlla la potenza del segnale: maggiore è, più facile è il compito dell'inferenza di recuperare $\mathbf{x}$ da $\mathbf{y}$.

Esiste un'identità generale chiamata **formula I-MMSE** [15] che collega la mutua informazione e l'MMSE per il problema del denoising:

$$\tfrac{d}{d\lambda}\tfrac{1}{p} I(\mathbf{x};\mathbf{y}) = \tfrac{1}{2}\mathbb{E}_{\mathbf{x},\mathbf{y}} \mathrm{MMSE}_p = \tfrac{1}{2}\mathbb{E}\,|\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{y}]\|^2.$$

Questa relazione è l'equivalente dell'identità termodinamica $f_p' = \langle m_p \rangle$ in (7), ma per inferenza ad alta-d (sotto l'assunto di rumore gaussiano).

Chiariamo la connessione tra entropia libera e mutua informazione: la log-funzione di partizione nella formula di Bayes si legge

$$\mathbb{E}_{\mathbf{y}} \ln \mathcal{Z}(\mathbf{y}) = \int d\mathbf{y} \mathbb{P}(\mathbf{y}) \ln \mathbb{P}(\mathbf{y}) = -H(\mathbf{y}).$$

Pertanto l'entropia libera attesa è collegata all'entropia di Shannon dei dati:

$$-\mathbb{E}_{\mathbf{y}} f_p(\mathbf{y}) = \tfrac{1}{p} H(\mathbf{y}).$$

nal is known, which is therefore the noise contribution. E.g., in a Gaussian **denoising model** $y = \sqrt{\lambda}\, x + z$ with $z \sim \mathcal{N}(0,1)$ we have $H(y \mid x) = H(z) = \ln(2\pi e)/2$. The mutual information is thus the information carried by the data purely about the signal. As such it quantifies the information-theoretic limits of inference, and computing it in high-d settings is a key goal of information theory. In the Gaussian denoising model it is an exercise to show that its explicit expression reads (here x^*, x are i.i.d. from $\mathbb{P}(x)$)

$$I(x;y) = \tfrac{\lambda}{2}\mathbb{E}[x^2] - \mathbb{E}_{x^*} \ln \mathbb{E}_x e^{\lambda x^* x + \sqrt{\lambda} z x - \frac{\lambda}{2} x^2}. \quad (9)$$

Denoising, the I-MMSE formula and the information-theoretic interpretation of the free entropy

Consider the general denoising model, where $\mathbf{x}$ can be a vector, matrix, etc: $\mathbf{y} = \sqrt{\lambda}\,\mathbf{x} + \mathbf{z}$. The r.v. $\mathbf{z}$ has same dimension as the signal and has i.i.d. standard normal entries. The SNR $\lambda > 0$ controls the signal strenght: the higher, the easier is the inference task of recovering $\mathbf{x}$ from $\mathbf{y}$.

There exists a general identity called **I-MMSE formula** [15] relating the mutual information and the MMSE for the denoising model:

$$\tfrac{d}{d\lambda}\tfrac{1}{p} I(\mathbf{x};\mathbf{y}) = \tfrac{1}{2}\mathbb{E}_{\mathbf{x},\mathbf{y}} \mathrm{MMSE}_p = \tfrac{1}{2}\mathbb{E}\|\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{y}]\|^2.$$

This relation is the equivalent of the thermodynamic identity $f_p' = \langle m_p \rangle$ in (7), but for high-d inference (under Gaussian noise).

Let us clarify the connection between free entropy and mutual information. The expected log-partition function in the Bayes formula reads

$$\mathbb{E}_{\mathbf{y}} \ln \mathcal{Z}(\mathbf{y}) = \int d\mathbf{y} \mathbb{P}(\mathbf{y}) \ln \mathbb{P}(\mathbf{y}) = -H(\mathbf{y}).$$

Therefore the expected free entropy is linked to the Shannon entropy of the data:

$$-\mathbb{E}_{\mathbf{y}} f_p(\mathbf{y}) = \tfrac{1}{p} H(\mathbf{y}).$$

Quindi le mutue informazioni verificano

$$\frac{1}{p}I(\mathbf{x};\mathbf{y}) = -\mathbb{E}f_p(\mathbf{y}) - \frac{1}{p}H(\mathbf{y} \mid \mathbf{x}). \qquad (10)$$

Il termine $\frac{1}{p}H(\mathbf{y} \mid \mathbf{x}) = \frac{1}{p}H(\mathbf{z}) = \frac{1}{2}\ln(2\pi e)$ è banale (perché il rumore ha componenti i.i.d.).

Un altro modo per vedere la connessione è partire dalla definizione termodinamica di entropia libera: l'entropia di Shannon della distribuzione di Gibbs-Boltzmann (la posterior) meno l'energia interna (ricordiamo che qui $\beta = 1$):

$$p\mathbb{E}f_p(\mathbf{y}) = H(\mathbf{x} \mid \mathbf{y}) - \mathbb{E}\langle\mathcal{H}(\mathbf{x};\mathbf{y})\rangle, \qquad (11)$$

dove $\mathcal{H}(\mathbf{x};\mathbf{y}) = -\ln\mathbb{P}(\mathbf{y} \mid \mathbf{x}) - \ln\mathbb{P}(\mathbf{x})$ è l'Hamiltoniana che definisce la posterior. Concentriamoci sul modello di denoising gaussiano: usando la formula di Bayes (1) l'energia interna verifica

$$\begin{aligned}\int d\mathbf{x}d\mathbf{y}\,\mathbb{P}(\mathbf{y})\mathbb{P}(\mathbf{x} \mid \mathbf{y})\mathcal{H}(\mathbf{x};\mathbf{y}) \\ = \int d\mathbf{x}d\mathbf{y}\,\mathbb{P}(\mathbf{x})\mathbb{P}(\mathbf{y} \mid \mathbf{x})\mathcal{H}(\mathbf{x};\mathbf{y}) \\ = \int d\mathbf{x}d\mathbf{z}\,\mathbb{P}(\mathbf{x})\mathbb{P}(\mathbf{z})\mathcal{H}(\mathbf{x};\sqrt{\lambda}\,\mathbf{x}+\mathbf{z})\end{aligned}$$

utilizzando il cambio di variabile $\mathbf{y} = \sqrt{\lambda}\,\mathbf{x} + \mathbf{z}$. Poiché il rumore è i.i.d. Gaussiano la probabilità $\mathbb{P}(\mathbf{y} \mid \mathbf{x})$ è una misura gaussiana multivariata con media $\sqrt{\lambda}\,\mathbf{x}$ e covarianza unitaria, e quindi $\mathbb{P}(\mathbf{z})$ è una gaussiana multivariata standard dopo il cambio di variabile. Perciò

$$\mathcal{H}(\mathbf{x};\mathbf{y}) = \tfrac{1}{2}\|\sqrt{\lambda}\,\mathbf{x} - \mathbf{y}\|^2 + \tfrac{p}{2}\ln(2\pi) - \ln\mathbb{P}(\mathbf{x}).$$

Finalmente raggiungiamo l'espressione per l'energia interna

$$\begin{aligned}\mathbb{E}\langle\mathcal{H}(\mathbf{x};\mathbf{y})\rangle &= \tfrac{1}{2}\mathbb{E}\|\mathbf{z}\|^2 + \tfrac{p}{2}\ln(2\pi) - \mathbb{E}\ln\mathbb{P}(\mathbf{x}) \\ &= \tfrac{p}{2}\ln(2\pi e) + H(\mathbf{x}).\end{aligned}$$

Usando questa, congiuntamente a (11), in $\frac{1}{p}I(\mathbf{x};\mathbf{y}) = \frac{1}{p}H(\mathbf{x}) - \frac{1}{p}H(\mathbf{x} \mid \mathbf{y})$ recuperiamo (10) e $-\mathbb{E}f_p = \frac{1}{p}H(\mathbf{y})$. Pertanto un fisico che cerca di calcolare l'entropia libera e un teorico dell'informazione che studia la mutua informazione stanno effettivamente mirando allo stesso obiettivo.

Grazie alla relazione I-MMSE il parametro d'ordine MMSE può essere derivato da $I(\mathbf{x};\mathbf{y})$, alla stregua della magnetizzazione derivabile dall'entropia libera nei modelli di meccanica statistica, almeno "in teoria ". Infatti, calcolare gli integrali p-dimensionali necessari per ottenere i potenziali termodinamici (mutua informazione, entropia

So the mutual information verifies

$$\frac{1}{p}I(\mathbf{x};\mathbf{y}) = -\mathbb{E}f_p(\mathbf{y}) - \frac{1}{p}H(\mathbf{y} \mid \mathbf{x}). \qquad (10)$$

The term $\frac{1}{p}H(\mathbf{y} \mid \mathbf{x}) = \frac{1}{p}H(\mathbf{z}) = \frac{1}{2}\ln(2\pi e)$ is trivial (because the noise has i.i.d. components).

Another way to see the connection is by starting from the thermodynamic definition of free entropy: the Shannon entropy of the Gibbs-Boltzmann distribution (the posterior) minus the internal energy (recall $\beta = 1$):

$$p\mathbb{E}f_p(\mathbf{y}) = H(\mathbf{x} \mid \mathbf{y}) - \mathbb{E}\langle\mathcal{H}(\mathbf{x};\mathbf{y})\rangle, \qquad (11)$$

where $\mathcal{H}(\mathbf{x};\mathbf{y}) = -\ln\mathbb{P}(\mathbf{y} \mid \mathbf{x}) - \ln\mathbb{P}(\mathbf{x})$ is the Hamiltonian defining the posterior. We focus on the Gaussian denoising model. Using the Bayes formula (1) the internal energy verifies

$$\begin{aligned}\int d\mathbf{x}d\mathbf{y}\,\mathbb{P}(\mathbf{y})\mathbb{P}(\mathbf{x} \mid \mathbf{y})\mathcal{H}(\mathbf{x};\mathbf{y}) \\ = \int d\mathbf{x}d\mathbf{y}\,\mathbb{P}(\mathbf{x})\mathbb{P}(\mathbf{y} \mid \mathbf{x})\mathcal{H}(\mathbf{x};\mathbf{y}) \\ = \int d\mathbf{x}d\mathbf{z}\,\mathbb{P}(\mathbf{x})\mathbb{P}(\mathbf{z})\mathcal{H}(\mathbf{x};\sqrt{\lambda}\,\mathbf{x}+\mathbf{z})\end{aligned}$$

using the change of variable $\mathbf{y} = \sqrt{\lambda}\,\mathbf{x} + \mathbf{z}$. As the noise is i.i.d. Gaussian the likelihood $\mathbb{P}(\mathbf{y} \mid \mathbf{x})$ is a multivariate Gaussian measure with mean $\sqrt{\lambda}\,\mathbf{x}$ and identity covariance, and thus $\mathbb{P}(\mathbf{z})$ is a standard multivariate Gaussian after the change of variable. Therefore

$$\mathcal{H}(\mathbf{x};\mathbf{y}) = \tfrac{1}{2}\|\sqrt{\lambda}\,\mathbf{x} - \mathbf{y}\|^2 + \tfrac{p}{2}\ln(2\pi) - \ln\mathbb{P}(\mathbf{x}).$$

We finally reach that the internal energy

$$\begin{aligned}\mathbb{E}\langle\mathcal{H}(\mathbf{x};\mathbf{y})\rangle &= \tfrac{1}{2}\mathbb{E}\|\mathbf{z}\|^2 + \tfrac{p}{2}\ln(2\pi) - \mathbb{E}\ln\mathbb{P}(\mathbf{x}) \\ &= \tfrac{p}{2}\ln(2\pi e) + H(\mathbf{x}).\end{aligned}$$

Using this as well as (11) in $\frac{1}{p}I(\mathbf{x};\mathbf{y}) = \frac{1}{p}H(\mathbf{x}) - \frac{1}{p}H(\mathbf{x} \mid \mathbf{y})$ we recover (10) and $-\mathbb{E}f_p = \frac{1}{p}H(\mathbf{y})$. Therefore a physicist trying to compute the free entropy and an information theorist the mutual information are actually aiming for the very same goal.

Thanks to the I-MMSE relation the MMSE order parameter can be derived from $I(\mathbf{x};\mathbf{y})$, or the magnetisation from the free entropy in statistical mechanics models, at least "in theory". Indeed, computing the p-dimensional integrals necessary to obtain the thermodynamic potentials (mutual information, free entropy) or the order parameters "directly" is generally a daunting task. But

libera) o i parametri d'ordine "direttamente" è generalmente un compito arduo. Ma come discuteremo verso la fine, in alcuni problemi ad alta-d, questo può essere ridotto a un problema di ottimizzazione scalare (molto) più semplice grazie al fenomeno della concentrazione in misura.

Consideriamo una prior fattorizzata, $\mathbb{P}(\mathbf{x}) = \prod_i P(x_i)$. In questo contesto, il problema del denoising è un "buon" esempio di problema di inferenza ad alta-d, con transizioni di fase e un ricco diagramma di fase? No. In effetti nel modello $\mathbf{y} = \sqrt{\lambda}\,\mathbf{x} + \mathbf{z}$ ogni punto dei dati $y_i(x_i, z_i)$ è solo funzione di un singolo segnale e componenti di rumore. Le v.c. $(x_i, z_i)_{i\leq p}$ sono i.i.d. per l'ipotesi di fattorizzazione. Di conseguenza, l'MMSE dell'intero segnale $\mathbf{x}$ è uguale all'MMSE di una singola voce in quanto sono statisticamente equivalenti: $\mathbb{E}\,\mathrm{MMSE}_p = \mathbb{E}[(\mathbb{E}[x_1 \mid y_1] - x_1)^2]$. Questa quantità è facilmente dimostrabile essere

$$\mathbb{E}[x^2] - \mathbb{E}_{z,x^*}\left[x^* \frac{\mathbb{E}_x\, x\, e^{-\frac{1}{2}(\sqrt{\lambda}(x^*-x)+z)^2}}{\mathbb{E}_x\, e^{-\frac{1}{2}(\sqrt{\lambda}(x^*-x)+z)^2}}\right] \qquad (12)$$

dove x, x^* sono i.i.d. da P e z è una v.c. gaussiana standard. Graficando questo parametro d'ordine MMSE in funzione del parametro di controllo λ, otteniamo una curva continua non-crescente, che svanisce al tendere di $\lambda \to \infty$: non così eccitante. Questo perché le variabili (x_i) sono in effetti **disaccoppiate** e il problema collassa in p problemi di inferenza scalare/a bassa dimensionalità equivalenti $y_i = \sqrt{\lambda}\, x_i + z_i$. E, siccome sono tutti statisticamente equivalenti, studiarne uno è sufficiente. Manca qualcosa nel problema per trasformare lo scenario in qualcosa di più ricco. Questo problema di denoising manca di un ingrediente chiave dei sistemi complessi: **correlazioni** tra gli ingressi del segnale indotte da interazioni non banali tra (x_i) nell'hamiltoniana.

Un paradigma per l'inferenza ad alta-d: il modello di spike di Wigner

Nell'inferenza ad alta risoluzione un modello importante è il **modello di spike di Wigner (SW)**, chiamato anche **fattorizzazione di matrici di rango basso**. Come vedremo, è un cugino stretto dei modelli Ising ed SK in meccanica statistica. È stato introdotto nella teoria delle matrici casuali

as we will discuss towards the end, in some high-d problems, this can be reduced to a (much) simpler scalar optimisation problem thanks to the concentration of measure phenomenon.

Consider a factorized prior $\mathbb{P}(\mathbf{x}) = \prod_i P(x_i)$. In this setting, is the denoising model a "good" example of high-d inference problem, with phase transitions and a rich phase diagram? No. Indeed in model $\mathbf{y} = \sqrt{\lambda}\,\mathbf{x} + \mathbf{z}$ each data point $y_i(x_i, z_i)$ is only function of a single signal and noise components. The r.vs. $(x_i, z_i)_{i\leq p}$ are i.i.d. by the factorization assumption. As a consequence the MMSE of the whole signal $\mathbf{x}$ equals the MMSE of a single entry as they are all statistically equivalent: $\mathbb{E}\,\mathrm{MMSE}_p = \mathbb{E}[(\mathbb{E}[x_1 \mid y_1] - x_1)^2]$. This quantity is easily shown to be

$$\mathbb{E}[x^2] - \mathbb{E}_{z,x^*}\left[x^* \frac{\mathbb{E}_x\, x\, e^{-\frac{1}{2}(\sqrt{\lambda}(x^*-x)+z)^2}}{\mathbb{E}_x\, e^{-\frac{1}{2}(\sqrt{\lambda}(x^*-x)+z)^2}}\right] \qquad (12)$$

where x, x^* are i.i.d. from P and z is a standard Gaussian r.v. Plotting this MMSE order parameter as a function of the λ control parameter, we get a smooth continuous non-increasing curve, that vanishes as $\lambda \to \infty$. Not so exciting. This is because the variables (x_i) are in fact **decoupled** and the problem collapses onto p parallel equivalent low-dimensional/scalar inference problems $y_i = \sqrt{\lambda}\, x_i + z_i$. And all are statistically equivalent so studying one is enough. Something is missing in the model in order to turn the picture into something richer. The denoising model lacks a key ingredient of complex systems: **correlations** among the signal entries induced by non-trivial interactions between the (x_i) in the Hamiltonian.

A paradigm of high-d inference: the spike Wigner model

In high-d inference an important model is the **spike Wigner (SW) model**, also called **low-rank matrix factorisation**. As we will discuss it is a close cousin of the Ising and SK models in statistical mechanics. It was introduced in random matrix theory as a simple model of principal components analaysis [16], which is the most widely

come semplice modello di analisi delle componenti principali [16], che è la tecnica di riduzione della dimensionalità maggiormente utilizzata.

Sia $\mathbf{z} = (z_{ij})_{i,j=1}^{n}$ una matrice di rumore con ingressi normali standard i.i.d. $z_{ij} \sim \mathcal{N}(0,1)$; questa è chiamata matrice di Wigner. Nel modello SW i dati sono (la parte triangolare superiore di) $\mathbb{R}^{p\times p} \ni \mathbf{y} = \sqrt{\lambda/p}\,\mathbf{x}\mathbf{x}^\intercal + \mathbf{z}$, o, per componenti,

$$y_{ij} = \sqrt{\frac{\lambda}{p}} x_i x_j + z_{ij} \quad \text{per} \quad 1 \le i < j \le p. \tag{13}$$

Il segnale $\mathbf{x}$ è una realizzazione della prior $\mathbb{P}(\mathbf{x}) = \prod_{i=1}^{p} P(x_i)$. Sfruttando il fatto che la verosimiglianza è la misura gaussiana multivariata, la posterior si legge (prescindendo da termini costanti che sono semplificati con la normalizzazione)

$$\begin{aligned}\mathbb{P}(\mathbf{x} \mid \mathbf{y}) = &\tfrac{1}{\mathcal{Z}(\mathbf{y})} \exp\Big\{ \textstyle\sum_{i=1}^{p} \ln P(x_i) \\ &- \tfrac{1}{2} \textstyle\sum_{i<j}^{p} (y_{ij} - \sqrt{\tfrac{\lambda}{p}}\, x_i x_j)^2 \Big\}.\end{aligned} \tag{14}$$

Ora vediamo la presenza di interazioni a coppie nell'Hamiltoniana, quindi le (x_i) non sono più disaccoppiati ed il problema non può essere ridotto a problemi indipendenti di inferenza scalare: questo sistema *è* complesso.

Si noti che le informazioni sul segno di $\mathbf{x}$ vengono perse per la simmetria $\pm\mathbf{x}$ in questa misura ogni volta che $P(x_i)$ è pari, ad esempio, quando si considera un segnale con ± 1 ingressi uniformi. In tali situazioni $\mathbb{P}(\mathbf{x} \mid \mathbf{y}) = \mathbb{P}(-\mathbf{x} \mid \mathbf{y})$ in modo che $\mathbb{E}[\mathbf{x} \mid \mathbf{y}] = (0)$. Quindi, in generale, ha più senso considerare la matrice di rango uno $\mathbf{x}\mathbf{x}^\intercal = (x_i x_j)_{i,j=1}^{p}$ come segnale nascosto (chiamato "picco ", i.e. *spike*). Ad ogni modo se lo statistico può recuperare lo spike, può accedere a $|\mathbf{x}|$ trovando il suo autovettore. Il rumore $\mathbf{z}$ rappresenta una fonte incontrollata di casualità che corrompe lo spike. Il compito dello statistico è quindi inferire $\mathbf{x}\mathbf{x}^\intercal$ nel modo più accurato possibile dato $\mathbf{y}$ e la conoscenza del processo di generazione dei dati (vale a dire il modello (13), ma non la realizzazione specifica di $\mathbf{x}$ né $\mathbf{z}$). Potremmo generalizzare ad altri tipi di rumore (non solo gaussiano né additivo), ma il quadro qualitativo non cambierebbe molto.

Il ridimensionamento $1/\sqrt{p}$ del RSR in (13) serve a rendere il compito dell'inferenza né impossibile né banale: qualsiasi altro ridimensionamento trasformerebbe il problema, nel limite di sistema di grandi dimensioni $p \to \infty$, in un problema con molto interesse. Per "non interessante" intendiamo

used dimensionality reduction technique.

Let $\mathbf{z} = (z_{ij})_{i,j=1}^{n}$ be a noise matrix with independent i.i.d. standard normal entries $z_{ij} \sim \mathcal{N}(0,1)$; this is called a Wigner matrix. In the SW model the data is (the upper triangular part of) $\mathbb{R}^{p\times p} \ni \mathbf{y} = \sqrt{\lambda/p}\,\mathbf{x}\mathbf{x}^\intercal + \mathbf{z}$, or componentwise,

$$y_{ij} = \sqrt{\frac{\lambda}{p}} x_i x_j + z_{ij} \quad \text{for} \quad 1 \le i < j \le p. \tag{13}$$

The signal $\mathbf{x}$ is a realisation of the prior $\mathbb{P}(\mathbf{x}) = \prod_{i=1}^{p} P(x_i)$. Using that the likelihood is the standard multivariate Gaussian measure the posterior reads (constant terms are simplified with the normalization)

$$\begin{aligned}\mathbb{P}(\mathbf{x} \mid \mathbf{y}) = &\tfrac{1}{\mathcal{Z}(\mathbf{y})} \exp\Big\{ \textstyle\sum_{i=1}^{p} \ln P(x_i) \\ &- \tfrac{1}{2} \textstyle\sum_{i<j}^{p} (y_{ij} - \sqrt{\tfrac{\lambda}{p}}\, x_i x_j)^2 \Big\}.\end{aligned} \tag{14}$$

Now we see pairwise interactions in the Hamiltonian, so the (x_i) are not anymore decoupled and the model cannot be reduced to independent scalar inference problems: this system *is* complex.

Note that the information about the sign of $\mathbf{x}$ is lost by $\pm\mathbf{x}$ symmetry in this measure whenever $P(x_i)$ is even, e.g., when considering a signal with ± 1 uniform entries. In such situations $\mathbb{P}(\mathbf{x} \mid \mathbf{y}) = \mathbb{P}(-\mathbf{x} \mid \mathbf{y})$ so that $\mathbb{E}[\mathbf{x} \mid \mathbf{y}] = (0)$. Therefore it makes more sense in general to consider the rank-one matrix $\mathbf{x}\mathbf{x}^\intercal = (x_i x_j)_{i,j=1}^{p}$ as hidden signal (called "spike"). Anyway if the statistician can recover the spike, it may access $|\mathbf{x}|$ by finding its eigenvector. The noise $\mathbf{z}$ represents a uncontrolled source of randomness that corrupts the spike. The statistician task is then to infer $\mathbf{x}\mathbf{x}^\intercal$ as accurately as possible given $\mathbf{y}$ and the knowledge of the data-generating process (namely the model (13), but not the specific realization of $\mathbf{x}$ nor $\mathbf{z}$). We could generalise to other type of noise (not only Gaussian nor additive), but the qualitative picture would not change much.

The scaling $1/\sqrt{p}$ of the SNR in (13) is there to make the inference task nor impossible nor trivial. Any other scaling would turn the problem, in the large-system limit $p \to \infty$, into a model with not much interest. By "uninteresting" we

blema di scarso interesse. Con "non interessante" si intende che lo spike (medio asintotico)-MMSE

$$\text{MMSE} := \lim_p \tfrac{1}{p^2}\mathbb{E}\|\mathbb{E}[\mathbf{x}\mathbf{x}^\intercal \mid \mathbf{y}] - \mathbf{x}\mathbf{x}^\intercal\|^2$$

sarebbe essenzialmente uguale a 0 per un ridimensionamento $p^{-\gamma} \gg 1/\sqrt{p}$, o al suo valore massimo per un ridimensionamento $p^{-\gamma} \ll 1/\sqrt{p}$, e questo indipendentemente da $\lambda = O(1)$. Qui $\|\cdot\|$ è la norma di Frobenius e $\mathbb{E}[\mathbf{x}\mathbf{x}^\intercal \mid \mathbf{y}] := \int d\mathbf{x}\, \mathbf{x}\mathbf{x}^\intercal\, \mathbb{P}(\mathbf{x} \mid \mathbf{y})$ è lo stimatore MMSE dello spike. Ma proprio per il ridimensionamento $\gamma = 1/2$ emerge un ricco **diagramma di fase** con **transizioni di teoria dell'informazione**.

Cerchiamo di capire più precisamente perché questo è il corretto ridimensionamento del RSR e, collegato a questo, che siamo davvero nel regime di alta-d.
Affinché il compito dell'inferenza non sia banale, dobbiamo collocarci nel regime ad alta-d. Come abbiamo già spiegato, ciò significa che il RSR totale per parametri, ovvero

$$\#\,\text{dati} \times \text{RSR}_\text{d} \div \#\,\text{parametri da inferire},$$

dovrebbe tendere ad una costante di ordine uno nel limite termodinamico. Abbiamo accesso a $p(p-1)/2$ punti di dati condizionatamente indipendenti $(y_{ij})_{i<j}$ e $\text{RSR}_\text{d} = \mathbb{E}[(\sqrt{\lambda/p}\, x_i x_j)^2] = (\mathbb{E}[x_1^2])^2\lambda/p$.
Lo verifichiamo

$$(\mathbb{E}[x_1^2])^2(\tfrac{p}{2}(p-1) \times \tfrac{\lambda}{p})\tfrac{1}{p} = (\mathbb{E}[x_1^2])^2\tfrac{\lambda}{2} + O(\tfrac{1}{p})$$

è effettivamente $O(1)$ come assumiamo $(\mathbb{E}[x_1])^2 = O(1)$. Questo spiega il ridimensionamento $1/\sqrt{p}$ nel modello (13): siamo nel regime di alta-d.

Esempi di applicazioni di questo modello sono (in tutti i casi che consideriamo la prior fattorizza come $\mathbb{P}(\mathbf{x}) = \prod_{i\le p} P(x_i)$):

- **Analisi delle componenti principali sparse:** Nel caso più semplice la prior $P = \text{Ber}(\rho)$ è Bernoulliana. Il compito è stimare la rappresentazione sparsa, nascosta, di basso rango $\mathbf{x}\mathbf{x}^\intercal$ di $\mathbf{y}$.
- **Identificazione di sotto-matrici:** Di nuovo $P = \text{Ber}(\rho)$. Si vuole qui estrarre sottomatrici di $\mathbf{y}$ di dimensione $\rho p \times \rho p$ con una media maggiore di quella dovuta al rumore di fondo; questo problema costituisce

mean that the (asymptotic average) spike-MMSE

$$\text{MMSE} := \lim_p \tfrac{1}{p^2}\mathbb{E}\|\mathbb{E}[\mathbf{x}\mathbf{x}^\intercal \mid \mathbf{y}] - \mathbf{x}\mathbf{x}^\intercal\|^2$$

would be essentially equal to 0 for a scaling $p^{-\gamma} \gg 1/\sqrt{p}$, or to its maximum value for a scaling $p^{-\gamma} \ll 1/\sqrt{p}$, and this independently of $\lambda = O(1)$. Here $\|\cdot\|$ is the Frobenius norm and $\mathbb{E}[\mathbf{x}\mathbf{x}^\intercal \mid \mathbf{y}] := \int d\mathbf{x}\, \mathbf{x}\mathbf{x}^\intercal\, \mathbb{P}(\mathbf{x} \mid \mathbf{y})$ is the MMSE estimator of the spike. But precisely for the scaling $\gamma = 1/2$ a rich **phase diagram** emerges with **information-theoretic phase transitions**.

Let us understand more precisely why this is the proper SNR scaling, and connected to that, that we are indeed in the high-d regime.
For the inference task not to be trivial we need place ourselves in the high-d regime. As we explained already this means that the total SNR per parameters, i.e.,

$$\#\,\text{data points} \times \text{SNR}_\text{d} \div \#\,\text{parameters to infer},$$

should tend to an order 1 constant in the thermodynamic limit. We have access to $p(p-1)/2$ conditionally independent data points $(y_{ij})_{i<j}$ and $\text{SNR}_\text{d} = \mathbb{E}[(\sqrt{\lambda/p}\, x_i x_j)^2] = (\mathbb{E}[x_1^2])^2\lambda/p$.
We verify that

$$(\mathbb{E}[x_1^2])^2(\tfrac{p}{2}(p-1) \times \tfrac{\lambda}{p})\tfrac{1}{p} = (\mathbb{E}[x_1^2])^2\tfrac{\lambda}{2} + O(\tfrac{1}{p})$$

is indeed $O(1)$ as we assume $(\mathbb{E}[x_1])^2 = O(1)$. This explains the scaling $1/\sqrt{p}$ in the observation model (13): we are in the high-d regime.

Examples of applications of this model are (we consider in all cases that the prior factorizes as $\mathbb{P}(\mathbf{x}) = \prod_{i\le p} P(x_i)$):

- **Sparse principal components analysis:** In the simplest case the prior $P = \text{Ber}(\rho)$ is Bernoulli. The task is to estimate the hidden sparse low-rank representation $\mathbf{x}\mathbf{x}^\intercal$ of $\mathbf{y}$.
- **Submatrix localization:** Again $P = \text{Ber}(\rho)$. One has then to extract a submatrix of $\mathbf{y}$ of size $\rho p \times \rho p$ with larger mean than the background noise matrix; this is an important model of hidden structure in computer

un importante modello di *struttura nascosta* nell'informatica.

- **Rilevamento di comunità in modelli a blocchi stocastici (SBM):** L'SBM (assortativo) è un modello di rete in cui sono più probabilmente osservati i bordi tra i nodi appartenenti alla stessa comunità. Dati questi margini osservati, il compito è inferire a quale comunità appartengono i nodi. Ad esempio, si suppone di conoscere la rete di amicizie in qualche social network. Nell'ipotesi che le persone che votano per lo stesso partito politico (tra due) siano collegate in questa rete con maggiore probabilità rispetto a coloro che votano contro, è possibile indovinare le due comunità di elettori (a meno di una permutazione globale)?

  Il recupero di due comunità di dimensione ρp e $(1-\rho)p$ in un SBM di p vertici è teoricamente "equivalente" al modello SW con una data prior (si veda [17] per il preciso significato dell'equivalenza)

$$P = \rho\delta_{\sqrt{(1-\rho)/\rho}} + (1-\rho)\delta_{-\sqrt{\rho/(1-\rho)}}. \tag{15}$$

- $\mathbb{Z}/2$**-Sincronizzazione:** La prior è Rademacher $P = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$. Il compito è dedurre gli stati dei nodi $\mathbf{x} \in \{-1,1\}^p$ (a meno di un segno globale) da prodotti rumorosi di coppie $\mathbf{y}$.

  Una possibile interpretazione: immaginiamo di poter chiedere alle coppie di individui (i,j) se sono d'accordo (+1) o meno (-1) su qualche domanda binaria "sì/no", ma che ci sia vietato chiedere ad ogni individuo i da solo qual sia la sua opinione sulla domanda e non fruiamo di idee a priori. Inoltre le risposte (y_{ij}) raccolte vengono trasmesse attraverso un canale di comunicazione (gaussiano) molto rumoroso. Ingenuamente, potremmo asserire che la coppia di individui (i,j) abbia la stessa opinione ogni volta che y_{ij} (uguale a $\sqrt{\lambda/p}\, x_i x_j + z_{ij}$, dove x_i è l'opinione dell'individuo i) è positivo, essendo z_{ij} centrato. Per le coppie tali che $y_{kl} < 0$ tenderemmo a concludere invece che $x_k x_l = -1$ (cioè che i singoli non sono d'accordo tra

science.

- **Community detection in the stochastic block model (SBM):** The (assortative) SBM is a network model where edges between nodes belonging to the same community are more probably observed. Given these observed edges, the task is to infer the community to which belong each nodes. For example, assume you know the network of friendships in some social network. Under the hypothesis that people voting for the same political party (among two) are connected in this network with higher probability than when they vote opposite parties, is it possible to guess the two communities of voters (up to a global permutation)?

  Recovering two communities of size ρp and $(1-\rho)p$ in a SBM of p vertices is information-theoretically "equivalent" to the SW model with prior (see [17] for the precise meaning of equivalence)

$$P = \rho\delta_{\sqrt{(1-\rho)/\rho}} + (1-\rho)\delta_{-\sqrt{\rho/(1-\rho)}}. \tag{15}$$

- $\mathbb{Z}/2$**Synchronization:** The prior is Rademacher $P = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$. The task is to infer the nodes states $\mathbf{x} \in \{-1,1\}^p$ (up to a global sign) from noisy pairwise products $\mathbf{y}$.

  A possible interpretation: imagine that you can ask to pairs (i,j) of individuals whether they agree (+1) or not (-1) on some binary "yes/no" question, but you cannot ask to any individual i alone what is her/his opinion on the question, and you have no a-priori idea about it. Moreover the answers (y_{ij}) you collect are transmitted through a very noisy (Gaussian) communication channel. Naively, you would naturally guess that the pair of individuals (i,j) have the same opinion whenever y_{ij} (equal to $\sqrt{\lambda/p}\, x_i x_j + z_{ij}$, where x_i is the opinion of individual i) is positive because z_{ij} is centered. For the pairs such that $y_{kl} < 0$ you would bet instead that $x_k x_l = -1$ (i.e., that they disagree). Of course with this naive approach contradictions will appear because of the noise. Let us say that you

loro). Ovviamente con questo approccio ingenuo appariranno delle contraddizioni a causa del rumore. Supponiamo di aver raccolto risposte molto rumorose (y_{ij}) per molte coppie (eventualmente tutte). E' possibile inferire in modo ottimale l'opinione di ogni individuo (a meno di un'inversione globale), ovvero chi sono gli "individui sincronizzati"? L'approccio ingenuo non è ottimale: quello che bisogna fare è usare la posterior (14) e calcolare lo stimatore MMSE (nel caso l'obiettivo sia minimizzare l'MSE) o lo stimatore MAP (se invece si vuole massimizzare la probabilità di trovare $\mathbf{x}$).

collected such noisy answers (y_{ij}) for many (all) pairs. Can you optimally infer the opinion of each individuals (up to global flip), i.e., who are the "synchronized individuals"? The naive approach is suboptimal. What one needs to do is to use the posterior (14) and compute the MMSE estimator (in case the goal is to minimize the MSE) or the MAP estimator (if instead one wants to maximize the probability of finding $\mathbf{x}$).

Una connessione con i modelli di Curie-Weiss e Sherrington-Kirkpatrick

Come promesso, ora stabiliamo una chiara connessione tra i modelli CW ed SK della meccanica statistica ed il modello SW dell'inferenza ad alta dimensionalità.

Consideriamo il caso binario $\mathbf{x} \in \{-1,1\}^p$ con prior di Rademacher: questo corrisponde al problema di $\mathbb{Z}/2$-sincronizzazione discusso in precedenza. In questa sezione sarà conveniente far apparire contemporaneamente sia il segnale originale (i.e. ground-truth signal) che la variabile che viene distribuita secondo la posterior. Rinomineremo quindi il segnale vero $\mathbf{x}^*$ dove $*$ sottolinea che è quello vero –che viene fissato quando si esegue l'inferenza– mentre $\mathbf{x}$ sono le variabili/spin che fluttuano secondo la posterior. Nel caso Rademacher la prior dà un contributo costante che si semplifica con la funzione di partizione e può quindi essere dimenticato nella posterior. Quindi, come si vede da (14), l'hamiltoniana del modello SW, quando esprime i dati in funzione del segnale e del rumore usando $y_{ij} = \sqrt{\lambda/p}\, x_i^* x_j^* + z_{ij}$ (e semplificando tutti i termini $\mathbf{x}$ indipendenti mediante la normalizzazione), si legge

$$\mathcal{H}_{\text{SW}}(\mathbf{x};\mathbf{y}) = -\sum_{i<j}^{p} \left(\tfrac{\lambda}{p} x_i^* x_j^* + \sqrt{\tfrac{\lambda}{p}} z_{ij}\right) x_i x_j$$

con $\mathbf{x} \in \{-1,1\}^p$. Questa è esattamente l'hamiltoniana SK (6) quando è presente solo il termine noise z_{ij} e λ è fisso ad uno. Il termine aggiuntivo correlato al segnale $-\sum_{i<j} \frac{\lambda}{p} x_i^* x_j^* x_i x_j$ è chiamato **planted term**, ed i modelli di inferenza so-

Link to the Curie-Weiss and Sherrington-Kirkpatrick models

As promised we now establish a clear connection between the CW and SK models from statistical mechanics and the SW model from high-d inference.

We consider the binary case $\mathbf{x} \in \{-1,1\}^p$ with Rademacher prior. This corresponds to the $\mathbb{Z}/2$-synchronization problem discussed above. In this section it will be convenient to make appear at the same time both the ground-truth signal and the variable that is distributed according to the posterior. Therefore we will rename the ground-truth signal $\mathbf{x}^*$ where the $*$ emphasizes that it is the true one, that is fixed when performing inference, while $\mathbf{x}$ are the variables/spins that fluctuate according to the posterior. In the Rademacher case the prior gives a constant contribution that simplifies with the partition function and can therefore be dropped in the posterior. Then, as seen from (14), the Hamiltonian of the SW model reads, when expressing the data as a function of the signal and noise using $y_{ij} = \sqrt{\lambda/p}\, x_i^* x_j^* + z_{ij}$ and simplifying all $\mathbf{x}$-independent terms with the normalization,

$$\mathcal{H}_{\text{SW}}(\mathbf{x};\mathbf{y}) = -\sum_{i<j}^{p} \left(\tfrac{\lambda}{p} x_i^* x_j^* + \sqrt{\tfrac{\lambda}{p}} z_{ij}\right) x_i x_j$$

with $\mathbf{x} \in \{-1,1\}^p$. This is exactly the SK Hamiltonian (6) when only the noise term z_{ij} is present and λ is set to one. The additional signal-related term $-\sum_{i<j} \frac{\lambda}{p} x_i^* x_j^* x_i x_j$ is called **planted term**, and inference models are **planted statistical mechanics models**. The planted term plays the role

no **modelli di meccanica statistica planted**. Il planted term svolge il ruolo di campo magnetico esterno che tende ad allineare gli spin nella direzione del segnale: trasporta le informazioni. Al contrario, il termine di rumore –che compete con quello planted– tende ad allineare gli spin in una direzione casuale non correlata al segnale. A seconda del valore del RSR λ che gioca un ruolo simile alla temperatura inversa β, un termine vince contro l'altro: per $\lambda > \lambda_c$ abbastanza alto vince il termine planted e gli spin "magnetizzano/polarizzano" nella direzione del segnale. Qui λ_c è la cosiddetta **soglia della teoria dell'informazione** (si veda la sezione successiva per maggiori dettagli). Questa polarizzazione è quantificata dall'overlap tra un campione $\mathbf{x}$ della posterior ed il segnale $\mathbf{x}^*$

$$m_p^* = \tfrac{1}{p}\sum_{i=1}^p x_i x_i^*. \tag{16}$$

Sia $m^* := \lim_p \mathbb{E}\langle m_p^*\rangle$. Scriviamo la media della posterior $\mathbb{E}[\cdot \mid \mathbf{y}]$ usando la notazione delle parentesi $\langle\cdot\rangle$ dalla meccanica statistica per enfatizzare che $\mathbb{P}(\mathbf{x} \mid \mathbf{y})$ è una distribuzione di Gibbs-Boltzmann; $\mathbb{E}$ media su tutte le variabili congelate $(\mathbf{x}^*, \mathbf{y})$ (o equivalentemente $(\mathbf{x}^*, \mathbf{z})$). Dopo alcune manipolazioni si può dimostrare che lo spike MMSE atteso si riferisce a questo parametro di ordine (sempre un'overlap)

$$\begin{aligned}\mathbb{E}\,\mathrm{MMSE}_p &= \tfrac{1}{p^2}\mathbb{E}\|\mathbf{x}^*(\mathbf{x}^*)^\intercal - \langle \mathbf{x}\mathbf{x}^\intercal\rangle\|^2 \\ &= (\mathbb{E}[(x_1^*)^2])^2 - \mathbb{E}\langle (m_p^*)^2\rangle.\end{aligned}$$

Come nel modello CW, la concentrazione in misura implica (nell'impostazione bayesiana ottimale): $m_p^* = m^* + o_p(1)$, ed a cascata la concentrazione dell'MMSE atteso (ed anche di quello non mediato in realtà) verso l'MMSE medio asintotico quando $p \to \infty$:

$$\mathrm{MMSE}_p \to (\mathbb{E}[(x_1^*)^2])^2 - (m^*)^2 =: \mathrm{MMSE}. \tag{17}$$

Nonostante l'hamiltoniana $\mathcal{H}_{\mathrm{SW}}$ somigli molto a quella dell'SK in quanto vi è disordine, la fenomenologia del modello SW è più vicina a quella del modello CW a causa del termine planted. Il parametro d'ordine m_p^* è la controparte nei problemi planted della magnetizzazione m_p nel modello CW, e si concentra, mentre l'overlap non lo fa nel modello SK: questo costituisce un'enorme differenza. Ciò complica drasticamente

of external magnetic field that tends to align the spins in the signal direction; it carries the information. In contrast the noise term, that competes with the planted one, tends to align the spins in a random direction that is uncorrelated with the signal. Depending on the value of the SNR λ that plays a similar role as the inverse temperature β, one term wins against the other: for high enough $\lambda > \lambda_c$ the planted term wins and the spins "magnetise/polarise" in the signal direction. Here λ_c is the so-called **information-theoretic threshold** (see next section for more details). This polarisation is quantified by the overlap between a sample $\mathbf{x}$ from the posterior and the signal $\mathbf{x}^*$

$$m_p^* = \tfrac{1}{p}\sum_{i=1}^p x_i x_i^*. \tag{16}$$

Let $m^* := \lim_p \mathbb{E}\langle m_p^*\rangle$. We write the posterior mean $\mathbb{E}[\cdot \mid \mathbf{y}]$ using the bracket notation $\langle\cdot\rangle$ from statistical mechanics to emphasize that $\mathbb{P}(\mathbf{x} \mid \mathbf{y})$ is a Gibbs-Boltzmann distribution; $\mathbb{E}$ is the average over all quenched variables $(\mathbf{x}^*, \mathbf{y})$ (or equivalently $(\mathbf{x}^*, \mathbf{z})$). After some manipulations one can demonstrate that the expected spike-MMSE relates to this overlap order parameter as

$$\begin{aligned}\mathbb{E}\,\mathrm{MMSE}_p &= \tfrac{1}{p^2}\mathbb{E}\|\mathbf{x}^*(\mathbf{x}^*)^\intercal - \langle \mathbf{x}\mathbf{x}^\intercal\rangle\|^2 \\ &= (\mathbb{E}[(x_1^*)^2])^2 - \mathbb{E}\langle (m_p^*)^2\rangle.\end{aligned}$$

As in the CW model the concentration of measure implies (in the Bayesian optimal setting): $m_p^* = m^* + o_p(1)$, and therefore concentration of the expected MMSE (and actually of the non-averaged one as well) towards the asymptotic average MMSE as $p \to \infty$:

$$\mathrm{MMSE}_p \to (\mathbb{E}[(x_1^*)^2])^2 - (m^*)^2 =: \mathrm{MMSE}. \tag{17}$$

Despite the Hamiltonian $\mathcal{H}_{\mathrm{SW}}$ ressembles a lot the one of the SK as there is disorder, the phenomenology of the SW model is closer to the one of the CW model due to the planted term. The order parameter m_p^* is the counterpart in planted problems of the magnetisation m_p in the CW model, and it concentrates, while the overlap does not in the SK model; that makes a huge difference. This complicates drastically the anal-

l'analisi del modello SK, vedere [7], e di altri modelli con **simmetria di replica rotta** [6, 12]. Questa è la terminologia della meccanica statistica per " mancanza di auto-media" dei parametri di ordine. Al contrario, i modelli CW e i modelli di inferenza ad alta-d nell'impostazione bayesiana ottimale sono **replica simmetrici**, cioè i loro parametri dell'ordine si concentrano sulla loro media come $p \to \infty$ [4, 5].

ysis of the SK model, see [7], and other models with **replica symmetry breaking** [6, 12]. This is the statistical mechanics terminology for "lack of self-averaging" of the order paramaters. In contrast the CW models and high-d inference models in the Bayesian optimal setting are **replica symmetric**, i.e., the order parameters do concentrate towards their mean as $p \to \infty$ [4, 5].

Transizioni di fase di teoria dell'informazione ed algoritmiche

Fino ad ora la nostra discussione è stata prevalentemente concettuale. Ma possiamo praticamente calcolare le principali grandezze ad alta-d che abbiamo introdotto (mutua informazione, entropia libera, MMSE) per comprendere e prevedere il comportamento degli algoritmi per problemi di inferenza ad alta-d? Continuiamo a concentrarci sul modello SW come esempio rappresentativo, ma la discussione seguente si applica in modo più generico.

Information-theoretic and algorithmic phase transitions

Until now our discussion was mostly conceptual. But can we practically compute the main high-d quantities we introduced (mutual information, free entropy, MMSE) in order to understand and predict the behavior of algorithms for high-d inference problems? We continue to focus on the SW model as a representative example, but the following discussion applies more generically.

"Single-letter formulas" per modelli di campo medio: la magia della concentrazione in misura

Derivare formule single-letter per quantità ad alta-d è spesso possibile per problemi appartenenti alla classe dei **modelli di campo medio**. Tali formule di solito si presentano sotto forma di un problema di ottimizzazione su una funzione di un parametro scalare. Nei modelli a campo medio ogni spin/variabile interagisce con molti altri, cioè con $O(p)$: parliamo in questo caso di un modello **denso**. Un'altra classe di modelli di campo medio sono i **modelli sparsi/diluiti**, dove la rete di interazioni tra variabili è tale che nel limite di $p \to \infty$, le variabili (x_i) interagiscono con un sottoinsieme casuale (finito) di $O(1)$ altre. Il modello SW è un modello a campo medio denso, poiché ogni variabile x_i interagisce con tutte le altre attraverso le interazioni di coppia $(\frac{1}{2}(y_{ij} - \sqrt{\lambda/p}\, x_i x_j)^2)_{j \le p}$. Per tali modelli esiste un arsenale di potenti metodi dalla meccanica statistica che sono in grado di ridurre la valutazione di quantità ad alta-d a problemi di ottimizzazione a bassa dimensione, in particolare il **metodo delle repliche** sviluppato nel contesto

"Single-letter formulas" for mean-field models: the magic of the concentration of measure

Deriving single-letter formulas for high-d quantities is often possible for problems belonging to the class of **mean-field models**. Such formulas usually come in the form of an optimization problem over a function of a scalar parameter. In mean-field models each spin/variable interacts with extensively many other ones, i.e., with $O(p)$: we speak in this case about a **dense** model. Another class of mean-field models are **sparse/dilute models**, where the network of interactions between variables is such that in the limit $p \to \infty$, variables (x_i) interact with a random subset of finitely many $O(1)$ other ones. The SW model is a dense mean-field model, as each variable x_i interacts with all the other ones through the pairwise interactions $(\frac{1}{2}(y_{ij} - \sqrt{\lambda/p}\, x_i x_j)^2)_{j \le p}$. For such models there exists an arsenal of powerful methods from statistical mechanics that are able to reduce the evaluation of high-d quantities to low-dimensional optimisation problems, in particular the **replica method** developed in the context of spin glasses [6, 12]. Such high-d

dei vetri di spin [6, 12]. Una tale riduzione da alta-d a bassa-d è un'altra bella manifestazione della concentrazione in misura.

Assumiamo ancora che la prior fattorizzi con $x_i \sim P$ i.i.d.: il metodo delle repliche (o il suo cugino stretto, il **metodo della cavità** [6, 12]) prevede che le informazioni reciproche per il modello SW verifichino come $p \to \infty$ (denotando $v := \mathbb{E}_P[x^2]$ e x^*, x sono i.i.d. generate da P, $z \sim \mathcal{N}(0,1)$ è una v.c. normale standard)

$$\tfrac{1}{p} I(\mathbf{x};\mathbf{y}) \to \min_{q\in[0,v]} \big\{ \tfrac{\lambda}{4}(qv)^2 + I(x; \sqrt{\lambda q}\, x + z) \big\}.$$

Qui $I(x; \sqrt{\lambda q}\, x + z)$ è la mutua informazione del modello di denoising gaussiano con RSR λq, dato da (9) che cambia λ in λq. Pertanto possiamo ottenere una formula effettiva per la mutua informazione. Equivale a 0 la derivata q della funzione $\{\cdots\}$ sopra - chiamata **potenziale replica simmetrico -**, il suo minimizzatore $q_{\min}$ verifica l'equazione di punto fisso

$$\tfrac{\lambda}{2}(q_{\min} - v) + \tfrac{d}{d\lambda} I(x; \sqrt{\lambda q}\, x + z)|_{q=q_{\min}} = 0.$$

Congiunto alla formula I-MMSE questo dà

$$q_{\min} = v - \mathrm{mmse}(x \mid \sqrt{\lambda q_{\min}}\, x + z) \qquad (18)$$

dove $\mathrm{mmse}(x \mid \sqrt{\lambda q_{\min}}\, x + z)$ è l'MMSE per il modello di denoising scalare; è dato da (12) con λ sostituito da $\lambda q_{\min}$. Ogni volta che è unico, è possibile dimostrare che il minimizzatore del potenziale replica simmetrico è uguale a $m^* := \lim_p \mathbb{E}\langle m_p^* \rangle$ (si ricordi (16)). Quindi da (17) otteniamo anche una "single-letter formula" per MMSE:

$$\mathrm{MMSE} = v^2 - q_{\min}^2. \qquad (19)$$

È assolutamente sorprendente che oggetti ad alta-d, che dipendono da così tante variabili casuali, possano essere ridotti a formule così semplici! Qui sta accadendo qualcosa di molto particolare: sia a livello di informazione reciproca che di MMSE appare il semplice modello scalare di denoising. L'analisi del modello SW ad alta-d collassa quindi sull'analisi di un problema di inferenza di una singola componente di segnale corrotta dal rumore gaussiano, con un RSR $\lambda q_{\min}$ dato da un'equazione di punto fisso non banale. Questa osservazione è generica per i modelli di

to low-d reduction is another beautiful manifestation of the concentration of measure.

Assume again that the prior factorizes with i.i.d. $x_i \sim P$. The replica method (or its close cousin the **cavity method** [6, 12]) predicts that the mutual information for the SW model verifies as $p \to \infty$ (denote $v := \mathbb{E}_P[x^2]$ and x^*, x are i.i.d. from P, $z \sim \mathcal{N}(0,1)$ is a standard normal r.v.)

$$\tfrac{1}{p} I(\mathbf{x};\mathbf{y}) \to \min_{q\in[0,v]} \big\{ \tfrac{\lambda}{4}(q-v)^2 + I(x; \sqrt{\lambda q}\, x + z) \big\}.$$

Here $I(x; \sqrt{\lambda q}\, x + z)$ is the mutual information of the Gaussian denoising model with SNR λq, given by (9) changing λ to λq. Therefore we can get an actual formula for the mutual information. Equating to 0 the q-derivative of the function $\{\cdots\}$ above –called **replica-symmetric potential**–, its minimizer $q_{\min}$ verifies the fixed point equation

$$\frac{\lambda}{2}(q_{\min} - v) + \frac{d}{d\lambda} I(x; \sqrt{\lambda q}\, x + z)|_{q=q_{\min}} = 0.$$

By the I-MMSE formula it gives

$$q_{\min} = v - \mathrm{mmse}(x \mid \sqrt{\lambda q_{\min}}\, x + z) \qquad (18)$$

where $\mathrm{mmse}(x \mid \sqrt{\lambda q_{\min}}\, x + z)$ is the MMSE for the scalar denoising model; it is given by (12) with λ replaced by $\lambda q_{\min}$. Whenever unique, the minimizer of the replica symmetric potential can be shown to be equal to $m^* := \lim_p \mathbb{E}\langle m_p^* \rangle$ (recall (16)). Therefore from (17) we also get a "single-letter formula" for the MMSE:

$$\mathrm{MMSE} = v^2 - q_{\min}^2. \qquad (19)$$

It is absolutely amazing that such high-d objects, that depend on so many random variables, can be reduced to such simple formulas! There is something very peculiar happening here: both at the level of the mutual information and of the MMSE the simple scalar denoising model appears. The analysis of the high-d SW model therefore collapses onto the analysis of an inference problem of a single signal component corrupted by Gaussian noise, with a SNR $\lambda q_{\min}$ given by a non-trivial fixed point equation. This obervation is generic for dense mean-fields models. For sparse problems things are a bit more subtle but essentially the same type of reduction

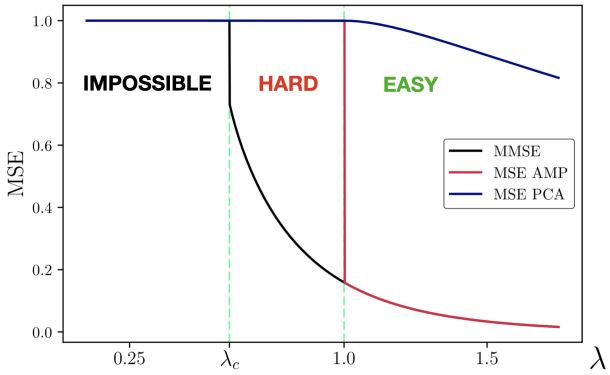


Figura 1: *Ita: Da [20]. Grafico dello spike-MMSE, il MSE dell'algoritmo AMP e della PCA naive per il modello SW con prior* (15) *con* $\rho = 0,05$. *Si osservano transizioni di fase del primo ordine sia di teoria dell'informazione che algoritmica. È presente un gap computazionale-statistico (fase difficile) tra le soglie critiche di teoria dell'informazione ed algoritmica.*
Eng: From [20]. Plot of the spike-MMSE, the MSE of the AMP algorithm and of naive PCA for the SW model with prior (15) *with* $\rho = 0.05$. *Information-theoretic and algorithmic first order phase transitions are observed. A computational-to-statistical gap (hard phase) is present between the information-theoretic and algorithmic thresholds.*

campo medio densi. Per i problemi sparsi le cose sono più sottili ma essenzialmente si verifica anche lì lo stesso tipo di riduzione di problemi da alta-d a bassa-d.

Si noti che a un certo punto abbiamo detto che il modello di denoising scalare non era così interessante in sé poiché non c'era transizione di fase nel suo MMSE. Ma qui anche se appare questo semplice modello, la complessità del SW si rivela nel fatto che le soluzioni dell'equazione di punto fisso (18) possono essere più di una. Quindi da un valore di RSR λ a uno vicino $\lambda + \varepsilon$, la soluzione $q_{\min}$ che minimizza il potenziale replica simmetrico (e quindi fornisce l'MMSE tramite (19)) può cambiare in modo discontinuo: si verifica quindi una transizione di fase.

Tutti questi risultati possono anche essere trasformati in affermazioni matematicamente rigorose. Complementari al metodo delle repliche, esistono i cosiddetti metodi **di cavità ed interpolante** [7, 26, 27, 28, 9], applicati al modello SW in [17]. Recentemente un'evoluzione del metodo di interpolazione per l'inferenza ad alta-d, chiamato **metodo di interpolazione adattiva**, ha avuto un grande successo nel dimostrare tali formule (comprese quelle fornite sopra per il modello SW) [29, 30, 24] [6]. Per coloro che sono interessati a saperne di più su queste tecniche di dimostrazione vedere [22, 20].

[6] Esiste anche un "approccio algoritmico " per dimostrare formule repliche simmetriche ad alta-d [18, 31].

from a high-d to low-d problems happens too.

Note that we said at some point that the scalar denoising model was not so interesting in itself as there was no phase transition in its MMSE. But here even if this simple model appears, the complexity of the SW is revealed in the fact that the solutions of the fixed point equation (18) may be more than one. So from one SNR value λ to a close one $\lambda + \varepsilon$, the solution $q_{\min}$ that minimizes the replica symmetric potential (and then gives the MMSE through (19)) may change discontinuously: a phase transition then occurs.

All these results can be even turned in mathematically rigorous statements. Complementary to the replica method, there exist the so-called **cavity and interpolation methods** [7, 26, 27, 28, 9], applied to the SW model in [17]. Recently an evolution of the interpolation method for high-d inference, called **adaptive interpolation method**, had great success in proving such fomulas (including the ones given above for the SW model) [29, 30, 24][6]. For those interested in knowing more about these proof techniques see [22, 20].

[6] There exists also an "algorithmic approach" to proving high-d replica symmetric formulas [18, 31].

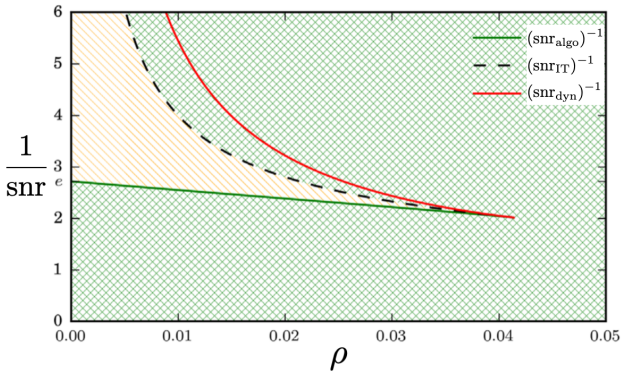


Figura 2: *Ita: Da [21]. Diagramma di fase del modello SW con parametri di Bernoulli $x_i \sim Ber(\rho)$ in funzione della scarsità ρ e del reciproco del RSR totale dato da RSR $:= \lambda\rho^2$. Non c'è transizione di fase nel sistema se $\rho > 0,0414$ e una transizione di fase del primo ordine altrimenti. La curva verde inferiore è la transizione di fase algoritmica dell'algoritmo AMP. La linea nera tratteggiata è la soglia critica di teoria dell'informazione. La zona tratteggiata in arancione è la regione difficile in cui il AMP non è ottimale (alla stregua di qualsiasi algoritmo di complessità sub-esponenziale noto). Nel resto del diagramma di fase (trattegiato in verde) il AMP fornisce nel limite di grandi dimensioni l'MMSE ottimale.*
Eng: From [21]. Phase diagram of the SW model with Bernoulli parameters $x_i \sim Ber(\rho)$ as a function of the sparsity ρ and inverse of the total SNR given by snr $:= \lambda\rho^2$. There is no phase transition in the system if $\rho > 0.0414$ and a first order phase transition else. The lower green curve is the algorithmic phase transition of the AMP algorithm. The dashed black line is the information theoretic threshold. The orange hashed zone is the hard region in which AMP is sub-optimal (as any known sub-exponential complexity algorithm). In the rest of the phase diagram (green hashed) the AMP provides in the large size limit the optimal MMSE.

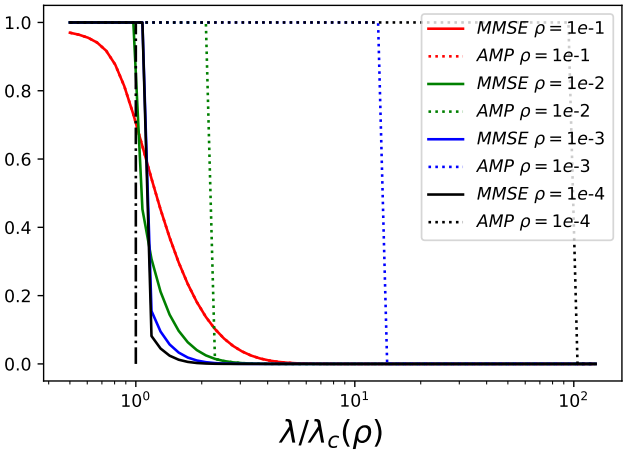


Figura 3: *Ita:Da [19]. Transizioni di fase in teoria dell'informazione del tipo tutto-o-niente e transizioni di fase algoritmiche ottenute mediante l'AMP per il modello SW con parametri di Bernoulli $x_i \sim Ber(\rho)$. Man mano che la scarsità ρ descresce, entrambe le transizioni diventano più nitide: una transizione tutto o niente appare nel limite $\rho \to 0$. L'asse orizzontale è su una scala logaritmica ed è relativo alla soglia critica in teoria dell'informazione $\lambda_c(\rho)$, essa stessa funzione di ρ (si veda [19] per la sua espressione). Il divario tra previsione statistica ed algoritmica diverge come $\rho \in 0$: dal punto di vista algoritmico diventa più difficile inferire il segnale.*
Eng: From [19]. All-or-nothing information-theoretic and AMP algorithmic phase transitions for the SW model with Bernoulli parameters $x_i \sim Ber(\rho)$. As the sparsity ρ dereases both transitions become sharper: an all-or-nothing transition appears in the limit $\rho \to 0$. Horizontal axis is on a log scale, and is relative to the information-theoretic threshold $\lambda_c(\rho)$, itself function of ρ (see [19] for its expression). The statistical-to-algorithmic gap diverges as $\rho \to 0$: it becomes algorithmically harder to infer the signal.

Dotati della formula esplicita (19) per MMSE siamo pronti ad esplorare il diagramma di fase del problema. Nella Figura 1 vengono tracciati sia l'MMSE che l'MSE raggiunto da due algoritmi per il modello SW. Questi algoritmi sono

Equipped with the explicit formula (19) for the MMSE we are ready to explore the phase diagram of the problem. In Figure 1 the MMSE as well as the MSE reached by two algorithms for the SW model is plotted. These algorithms

l'**analisi delle componenti principali (PCA)** e l'algoritmo di **trasmissione di messaggi approssimati (AMP)**. Nella PCA si calcola l'autovettore della matrice dei dati $\mathbf{y}$ associato all'autovalore massimo; questo è lo stimatore del segnale. Al di sopra di una certa **soglia algoritmica** questo stimatore di autovettori inizia ad allinearsi con il segnale in modo che il MSE si abbassi. Non discuteremo l'algoritmo AMP, ma essenzialmente ciò che conta è che molti ipotizzano che sia ottimale tra tutti gli algoritmi a bassa complessità, pratici in un'ampia classe di problemi di inferenza ad alta-d. Qui osserviamo infatti che l'AMP richiede un RSR inferiore rispetto alla PCA per funzionare bene (cioè, può funzionare meglio a livelli di rumore più elevati). E quando funziona, offew prestazioni pari a quelle dello stimatore MMSE. Inoltre la sua prestazione, nel limite di $p \to \infty$ può essere rigorosamente prevista. Questo permette di ottenere le curve presentate qui, si veda [11, 21, 23, 24] per i dettagli.

Quello che osserviamo è uno scenario generico in inferenza ad alta-d con due tipi di transizioni di fase che delimitano tre fasi: $i)$ la **fase impossibile** è il regime in cui anche lo stimatore MMSE ottimale si comporta male (non meglio di ipotesi casuale). Pertanto è teoricamente impossibile inferire qualcosa sul segnale meglio di un'ipotesi casuale: non è un problema di calcolo, semplicemente non ci sono abbastanza informazioni. La soglia critica per questo è indicata con λ_c. Il regime in cui il RSR $\lambda \in (\lambda_c, \lambda_{\text{algo}})$ (dove in questo problema la soglia algoritmica $\lambda_{\text{algo}} = 1$ è la stessa per la PCA ed il AMP) è la **fase difficile**. Difficile nel senso algoritmico: significa che non conosciamo alcun algoritmo computazionalmente efficiente in grado di eguagliare le prestazioni dello stimatore MMSE ottimale. Infine $\lambda > \lambda_{\text{algo}}$ corrisponde alla **fase facile**: in questo regime conosciamo un algoritmo computazionalmente efficiente (AMP) in grado di eguagliare l'MMSE. In questo modello con questa specifica prior sia la transizione di teoria dell'informazione che quella algoritmica dell'AMP sono brusche/discontinue: sono del primo tipo di ordine. A volte sono continue come hre per la stima della PCA. La presenza di una fase difficile definisce un cosiddetto **divario computazionale-statistico** (un altro nome per il regime hard), e capire se tale gap sia fondamentale o meno è una delle principali questioni

are **principal component analysis (PCA)** and the **approximate message-passing (AMP)** algorithm. In PCA one computes the eigenvector of the data matrix $\mathbf{y}$ associated with the maximum eigenvalue; this is the estimator of the signal. Above some **algorithmic threshold** this eigenvector estimator starts to align with the signal so that the MSE lowers down. We will not discuss the AMP algorithm, but essentially what matters is that it is conjectured by many to be optimal among all low-complexity/practical algorithms in a broad class of high-d inference problems. Here we indeed observe that AMP requires a lower SNR than PCA to perform well (i.e., can perform better at higher noise levels). And when it works it matches the MMSE estimator performance. Moreover its performance in the limit $p \to \infty$ can be rigorously predicted. This allows to get the curves presented here, see [11, 21, 23, 24] for details.

What we observe is a generic scenario in high-d inference with two types of phase transitions delimiting three phases: $i)$ the **impossible phase** is the regime where even the optimal MMSE estimator performs poorly (not better than random guessing). Therefore it is information-theoretically impossible to infer anything about the signal better than random guessing. It is not a computational issue, there is simply not enough information. The information-theoretic threshold is denoted λ_c. The SNR regime $\lambda \in (\lambda_c, \lambda_{\text{algo}})$ (where in this problem the algorithmic threshold $\lambda_{\text{algo}} = 1$ is the same for PCA and AMP) is the **hard phase**. Hard is in the algorithmic sense. It means that we do not know any computationally efficient algorithm able to match the performance of the optimal MMSE estimator. Finally $\lambda > \lambda_{\text{algo}}$ corresponds to the **easy phase**: in this regime we do know a computationally efficient algorithm (AMP) able to match the MMSE. In this model with this specific prior both the information-theoretic and algorithmic transition of AMP are sharp/discontinuous: they are of the first order type. Sometimes they are continuous like hre for the PCA estimate. The presence of an hard phase defines a so-called **computational-to-statistical gap** (another name for the hard regime), and understanding whether such gap is fundamental or not is one of the main open question in the field. By fundamental we

aperte nel campo. Per fondamentale si intende se esiste effettivamente o meno in questa regione un algoritmo che performi in un tempo polinomiale (in p) in grado di battere il AMP e corrispondere all'MMSE.

Questi tre regimi sono separati da transizioni di fase. Consideriamo il modello SW con prior di Bernoulli di media ρ. Mostriamo le linee di transizione di fase nel piano $(1/(\lambda\rho^2), \rho)$ (questi sono i parametri di controllo; il RSR $= \lambda\rho^2$ è il naturale parametro RSR) nella figura 2. Prevedere l'andamento degli stimatori MMSE e AMP in ogni punto, permette di disegnare il diagramma di fase del problema. Osserviamo ampie regioni in verde dove il AMP è ottimale e la fase hard in arancione. Questo è simile al diagramma di fase dell'acqua nel piano (temperatura, pressione) con le fasi solida, liquida e gassosa. Questo tipo di immagini permette di leggere le limitazioni fondamentali e algoritmiche della ricostruzione del segnale al variare dei parametri di controllo.

Citiamo un'altra osservazione interessante. È stato recentemente dimostrato in [19] (basato sulle congetture di [21]) che le **transizioni di fase tutto o niente** avvengono in un regime di sparsità molto elevata $\rho \to 0$ (sempre considerando una prior di Bernoulli per gli ingressi del segnale). Ciò significa che, come osservato nella Figura 3, le transizioni diventano tanto nitide quanto possono essere in questo particolare limite. Ciò significa che quando la **dimensione effettiva del segnale** è molto più piccola della sua dimensione ambientale p, il segnale può essere o perfettamente dedotto, oppure per niente. Non vi è alcun crossover tra questi due comportamenti come si evince dalla Figura 1 ottenuta con una diluizione finita ρ. Qui la dimensione effettiva del segnale è ρp, cioè il numero atteso di componenti diverse da zero: scompare se confrontato con la dimensione ambientale p come $\rho \to 0$. Questa fenomenologia sembra molto generica e si verifica in un'ampia classe di altri modelli di inferenza ad alta-d [32]. Si ritiene che il successo della moderna elaborazione del segnale e dell'apprendimento automatico nei regimi ad alta dimensionalità sia in parte dovuto alla struttura dei dati stessi e al fatto che, anche se ad alta-d, hanno una dimensionalità effettiva inferiore, che viene poi sfruttata dagli algoritmi. Pertanto la progettazione e l'analisi di modelli semplici che siano

mean whether there actually exists or not in this region a polynomial-time (in p) algorithm able to beat AMP and match the MMSE.

These three regimes are separated by phase transitions. Consider the SW model with Bernoulli prior of mean ρ. We show the phase transitions lines in the $(1/(\lambda\rho^2), \rho)$ plane (these are the control parameters; snr $= \lambda\rho^2$ is the natural SNR parameter) in Figure 2. Predicting the performance of the MMSE and AMP estimators at each point, it allows to draw the phase diagram of the problem. We observe large regions in green where AMP is optimal, and the hard phase in orange. This is similar to the phase diagram of water in the (temperature, pressure) plane with the solid, liquid and gas phases. This kind of pictures allow to read fundamental and algorithmic limitations of signal reconstruction as control parameters are varied.

Let us mention another interesting observation. It was proven recently in [19] (based on conjectures in [21]) that **all-or-nothing phase transitions** happen in the regime of very high sparsity $\rho \to 0$ (still considering a Bernoulli prior for the signal entries). This means that, as observed in Figure 3, the transitions become as sharp as they can be in this particular limit. It means that when the **effective dimension of the signal** is much smaller than its ambient dimension p, the signal can be or perfectly infered, or not at all. There is no crossover between these two behaviors like in Figure 1 which is for a finite sparsity ρ. Here the effective dimension of the signal is ρp, i.e., the expected number of non-zero components. It vanishes when compared to the ambient dimension p as $\rho \to 0$. This phenomenology seems very generic and happens in a broad class of other high-d inference models [32]. The success of modern signal processing and machine learning in high-d regimes is believed to be partly due to the structure of the data itself and the fact that even if high-dimensional, it has lower effective dimensionality, that is then exploited by algorithms. Therefore designing and analysing simple models that are tractable and serve as idealized paradigms for this setting is of fundamental interest.

trattabili e fungano da paradigmi idealizzati per questo contesto è di fondamentale interesse.

Considerazioni conclusive

Abbiamo discusso del regime moderno delle statistiche ad alta-d. Concentrandoci sul modello di spike di Wigner come paradigma di inferenza ad alta dimensionalità, abbiamo dimostrato che l'inferenza può essere riformulata nel linguaggio della meccanica statistica. Come in modelli più fisici come i sistemi spin (e praticamente qualsiasi sistema sufficientemente complesso) il modello SW ha transizioni di fase che separano diversi regimi algoritmici di inferenza.

Per motivi di pedagogia ci siamo concentrati sul modello SW, ma gran parte dei concetti che abbiamo introdotto, la fenomenologia che abbiamo presentato e le conclusioni che abbiamo tratto sono molto più generali e si applicano a una classe estremamente ampia di problemi di inferenza ed apprendimento automatico. Per avere una visione più ampia e conoscere molti altri esempi di modelli di inferenza ad alta-d che possono essere trattati utilizzando l'approccio della meccanica statistica, raccomando l'eccellente review [11]. Si veda anche l'articolo [24]. Per i lettori matematicamente orientati può essere stimolante la lettura di [22] e [20]. I riferimenti classici sono i libri [33, 34].

Concluding remarks

We discussed the modern regime of high-d statistics. Focusing on the spike Wigner model as paradigm of high-d inference, we have shown that inference can be recast in the statistical mechanics language. As in more physical models like spins systems (and virtually any sufficiently complex system) the SW model has phase transitions separating different algorithmic regimes of inference.

For the sake of pedagogy we focused on the SW model. But a large part of the concepts we introduced, the phenomenology we presented and the conlusions we drew are much more general and apply to an extremely large class of inference and learning problems. In order to get a broader view and know about many more examples of high-d inference models that can be treated using the statistical mechanics approach I recommend the excellent review [11]. See also the article [24]. For mathematically oriented readers see [22] and [20]. Classical references are the books [33, 34].

[1] L. Wasserman: *All of statistics: a concise course in statistical inference*, Springer Science & Business Media, Berlin (2013).

[2] D. MacKay: *Information theory, inference and learning algorithms*, Cambridge Univ. Press, Cambridge (2003) http://www.inference.org.uk/mackay/itila/book.html

[3] E. J. Candès, M. B. Wakin: *An introduction to compressive sampling*, IEEE Sign. Proc. Mag. (2008), https://authors.library.caltech.edu/10092/

[4] J. Barbier: *Overlap matrix concentration in optimal Bayesian inference*, Information and Inference: a Journal of the IMA & arXiv preprint arXiv:1904.02808, (2020)

[5] J. Barbier, D. Panchenko: *Strong replica symmetry in high-dimensional optimal Bayesian inference*, arXiv preprint arXiv:2005.03115 (2020), https://arxiv.org/abs/2005.03115

[6] M. Mézard, A. Montanari: *Information, physics, and computation*, Oxford Univ. Press, Oxford (2009) https://web.stanford.edu/~montanar/RESEARCH/book.html

[7] D. Panchenko: *The Sherrington-Kirkpatrick model*, Springer Science & Business Media, Berlin (2013)

[8] M. Talagrand: *The Parisi formula,* Ann. of Math., 163 (2006) 221.

[9] F. Guerra, *Broken replica symmetry bounds in the mean field spin glass model*, Comm. Math. Phys. 233(1), 1-12, (2003)

[10] G. Parisi: *A sequence of approximated solutions to the Sherrington-Kirkpatrick model for spin glasses,* J. Phys. A, 13 (1980) L115.

[11] L. Zdeborová, F. Krzakala: *Statistical physics of inference: thresholds and algorithms,* Adv. in Phys., 65 (2016) 453.

[12] M. Mézard, G. Parisi, M .Virasoro: *Spin glass theory and beyond: an introduction to the replica method and its applications*, World Scientific Publishing, Singapore (1987).

[13] C. E. Shannon: *A mathematical theory of communication*, The Bell System Technical Journal, 27 (1948) 623.

[14] T. M. Cover, J. M. Thomas: *Elements of information theory*, John Wiley & Sons, New York (1999).

[15] D. Guo, S. Shamai, S. Verdú: *Mutual information and minimum mean-square error in Gaussian channels*, IEEE Trans. on Inform. Th., 5 (2005) 1261.

[16] I. M. Johnstone: *On the distribution of the largest eigenvalue in principal components analysis*, Ann. of Stat., 29 (2001) 295.

[17] M. Lelarge, L. Miolane: *Fundamental limits of symmetric low-rank matrix estimation*, Prob. Th. Rel. Fiel., 173 (2019) 859.

[18] M. Dia, J. Barbier, N. Macris, F. Krzakala, T. Lesieur, L. Zdeborová: *Mutual information for symmetric rank-one matrix estimation: a proof of the replica formula*, Adv. Neur. Inf. Proc. Sys., 29 (2016) 424.

[19] J. Barbier, N. Macris, C. Rush: *All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation*, Adv. Neur. Inf. Proc. Sys. & arXiv preprint arXiv:2006.07971, (2020)

[20] L. Miolane: *Fundamental limits of inference: A statistical physics approach*, PhD thesis (2020) https://hal.archives-ouvertes.fr/tel-02446988

[21] T. Lesieur, F. Krzakala, L. Zdeborová, *Constrained low-rank matrix estimation: phase transitions, approximate message passing and applications*, J. Stat. Mech. 7 (2017) 073403.

[22] J. Barbier: *Mean-field theory of high-dimensional Bayesian inference*, Course given at the school "Mathematical and Computational Aspects of Machine Learning", Scuola Normale Superiore di Pisa (2020) https://www.overleaf.com/read/yhsncssvbcqr

[23] F. Krzakala, M. Mézard, F .Sausset, Y. Sun, L. Zdeborová, *Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices*, J. Stat. Mech., 8 (2012) P08009.

[24] J. Barbier, F. Krzakala, N. Macris, L. Miolane, L. Zdeborová, *Optimal errors and phase transitions in high-dimensional generalized linear models*, Proc. Natl. Acad. Sci. USA ,116 (2019) 5451.

[25] J. Barbier, *Phase transitions: from physics to computer science*, Online "Basic Notions Seminar" from the ICTP Mathematics Department (2020) https://www.youtube.com/watch?v=q1VO5dmymFM&t=3077s&ab_channel=ICTPMathematics

[26] M. Talagrand: *Mean field models for spin glasses: volume I: basic examples*, Springer Science & Business Media, Berlin (2010).

[27] M. Talagrand: *Mean field models for spin glasses: volume II: advanced replica-symmetry and low temperature*, Springer Science & Business Media, Berlin (2010)

[28] F. Guerra, F. L. Toninelli: *The thermodynamic limit in mean field spin glass models*, Comm. Math. Phys., 230 (2002) 71.

[29] J. Barbier, N. Macris: *The adaptive interpolation method: a simple scheme to prove replica formulas in Bayesian inference*, Prob. Th. Rel. Fiel.,174 (2019) 1133.

[30] J. Barbier, N. Macris: *The adaptive interpolation method for proving replica formulas. Applications to the Curie–Weiss and Wigner spike models*, J. Phys. A 52, (2019) 294002.

[31] J. Barbier, N. Macris, M. Dia and F. Krzakala, *Mutual information and optimality of approximate message-passing in random linear estimation*, IEEE Trans. Inform. Th. (2020)

[32] C. Luneau, J. Barbier, N. Macris, *Information theoretic limits of learning a sparse rule*, Adv. Neur. Inf. Proc. Sys. & arXiv preprint arXiv:2006.11313, (2020)

[33] A. Engel, C. Van den Broecktitle, *Statistical mechanics of learning*, Cambridge University Press, Cambridge (2001)

[34] H. Nishimori, *Statistical physics of spin glasses and information processing: an introduction*, Clarendon Press, Oxford (2001).

Jean Barbier: è Assistant Professor all'Abdus Salam International Center for Theoretical Physics in Trieste, Italia. I suoi interessi di ricerca principali sono probabilità e statistica ad alta dimensionalità, teoria dell'informazione, meccanica statistica dei sistemi disordinati e sue connessioni interdisciplinary con inferenza, machine learning e computer science.

Jean Barbier: is an Assistant Professor at the Abdus Salam International Center for Theoretical Physics in Trieste, Italy. His main interests are in high-dimensional probability and statistics, information theory, statistical mechanics of disordered systems and its interdisciplinary connections with inference, machine learning and computer science.

Metodi di massima entropia

Michele Castellana *Laboratoire Physico–Chimie Curie, Institut Curie, PSL Research University; Sorbonne Universités, UPMC Univ. Paris 06*

I metodi di massima entropia (MME) costituiscono uno strumento teorico sistematico per costruire modelli per sistemi fisici: questi modelli devono essere consistenti con un insieme di misure, ma allo stesso tempo avere quanta meno struttura possibile. Questo metodo costituisce, in linea di principio, una strategia priva di *bias* per costruire un modello del fenomeno in considerazione, persino in presenza di una quantità limitata di dati. In quanto segue presenteremo una breve introduzione ai principali aspetti matematici e concettuali dei MME. La seconda parte di questo articolo sarà focalizzata sui possibili punti deboli, sia concettuali che tecnici, dei MME, in modo da fornire al lettore un'analisi critica di questi metodi.

Maximum-entropy methods (MEMs) provide a systematic theoretical framework for building models of physical systems which are consistent with some set of measurements, but otherwise have as little structure as possible. This constitutes, in principle, a bias-free, sensible strategy which allows one to model the phenomenon under consideration, even in the presence of a limited amount of data. In what follows, we will present a short introduction to the main mathematical and conceptual aspects of MEMs. The second part of this review will focus on the potential issues, both conceptual and technical, to which MEMs are subject, so as to provide a critical assessment of the method for the general reader.

Introduzione

Il lavoro pionierisitco di Shannon [1] ha dimostrato che l'entropia di una distribuzione di probabilità può indicare la quantità di `struttura` contenuta nella distribuzione. Shannon considerò un insieme di eventi $i = 1,...,N$ che avvengono con probabilità p_i ed un insieme di criteri intuitivi e minimali per una funzione $S[\boldsymbol{p}]$ che possa rappresentare la quantità di struttura contenuta in $\boldsymbol{p}$, in modo tale che più grande è S, minore è la quantità di struttura. Shannon dimostrò che l'unica quantità che soddisfa questi criteri è

Introduction

The pioneering work by Shannon [1] showed that the entropy of a probability distribution can be regarded as an indicator of the amount of 'structure' contained in such distribution. Shannon considered a set of events $i = 1,...,N$ which happen with probability p_i, and a set of intuitive, minimal criteria for a function $S[\boldsymbol{p}]$ which is required to represent the amount of structure contained in $\boldsymbol{p}$, i.e., the larger S, the smaller the structure. Shannon demonstrated that the only

l'entropia

$$S[\boldsymbol{p}] \equiv -\sum_i p_i \log p_i. \tag{1}$$

In particolare, i criteri per S sono i seguenti:

1. S è una funzione continua di $\boldsymbol{p}$.
2. Se tutti i p_i sono uguali, S è una funzione monotona crescente del numero di eventi N
3. Se un evento j è stato suddiviso in più eventi secondari, l'entropia $S[\boldsymbol{p}]$ originale è uguale alla somma di $S[\boldsymbol{p}']$, dove $\boldsymbol{p}'$ corrisponde alla probabilità degli eventi suddivisi, e dell'entropia degli eventi secondari, dove quest'ultima è pesata con la probabilità p'_j—vedi Fig. 1 per un esempio illustrativo.

I tre criteri di cui sopra possono essere interpretati nel modo seguente. Oltre alla condizione di continuità 1, il criterio 2 riflette il fatto che, se tutti gli eventi si verificano con la stessa probabilità, maggiore è il numero di eventi, N, minore è la quantità di struttura codificata in $\boldsymbol{p}$. Ad esempio, per $N = 1$, $\boldsymbol{p}$ sarebbe una distribuzione altamente strutturata, poiché il processo casuale in esame produrrebbe un singolo evento in modo deterministico. D'altra parte, per N grande, $\boldsymbol{p}$ assegnerebbe la stessa probabilità ad un gran numero di eventi, contenendo quindi una minore quantità di struttura ed essendo quindi meno informativa sull'esito del processo casuale.
Infine, la condizione 3 riflette l'idea che, se uno degli eventi è suddiviso in un insieme di eventi secondari, la quantità di struttura, o entropia, di $\boldsymbol{p}$ deve essere data dalla somma dell'entropia relativo agli eventi, più l'entropia degli eventi secondari.

Il metodo

La definizione quantitativa di cui sopra della struttura contenuta in una distribuzione di probabilità ha consentito diversi importanti progressi nel campo dell'inferenza statistica. Infatti, la relazione (1) ha consentito un'implementazione matematica diretta del principio del rasoio di Ockam [2]. Conosciuto anche come legge della parsimonia, il rasoio di Ockham è un principio che, in generale, afferma che la soluzione più semplice ad un problema è probabilmente quella corretta. A questo proposito, il termine `rasoio` si

quantity which satisfies these criteria is the entropy

$$S[\boldsymbol{p}] \equiv -\sum_i p_i \log p_i. \tag{1}$$

In particular, the criteria for S are the following:

1. S is a continuous function of $\boldsymbol{p}$.
2. If all p_is are equal, then S is a monotonically increasing function of the number of events. N
3. If an event j were broken into multiple subevents, then the original entropy $S[\boldsymbol{p}]$ equals the sum of $S[\boldsymbol{p}]$, where $\boldsymbol{p}'$ corresponds to the probability of the broken events, and the entropy of the subevents, where the latter is weighted with the probability p'_j—see Fig. 1 for an illustrative example.

The three criteria above allow for the following interpretation. In addition to the continuity condition 1, criterion 2 reflects the expectation that, if all events occur with the same probability, the larger the number of events, N, the smaller the amount of structure encoded in $\boldsymbol{p}$. For example, for $N = 1$, $\boldsymbol{p}$ would be a highly structured distribution, implying that the random process under consideration yields a single event in a deterministic way. On the other hand, for larger N, $\boldsymbol{p}$ would assign the same probability to a large number of events, thus being less informative, i.e., bearing a smaller amount of structure, on the outcome of the random process.
Finally, condition 3 reflects the idea that, if one of the events is broken into a set of subevents, then the amount of structure, or entropy, of $\boldsymbol{p}$ must be given by the sum of the entropy relative to the events, plus the entropy of the subevents.

The method

A quantitative definition for the amount of structure encoded in a probability distribution allowed for several important advances in the field of statistical inference. In fact, the relation (1) allowed for a direct, mathematical implementation of the long-standing principle of Ockam's razor [2]. Also known as the law of parsimony, Ockham's razor is a principle which, generally speaking, states that the simplest possible solution to a problem is most likely the correct one. In this regard, the term 'razor' refers to the act of

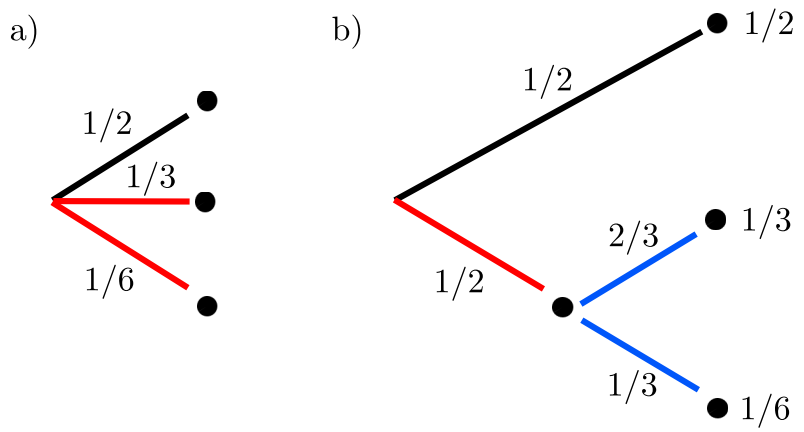


Figura 1: *Uno dei requisiti di Shannon per la funzione entropia. a) Tre eventi (punti) con probabilità* $\boldsymbol{p} = (1/2, 1/3, 1/6)$ *ed entropia* $S[\boldsymbol{p}]$*. b) Gli eventi in a) evidenziati in rosso sono sostituiti da un evento con probabilità* $1/2$*, suddiviso in due eventi secondari con probabilità* $2/3$ *e* $1/3$*. Complessivamente, in b) ci sono tre eventi possibili, che si verificano con probabilità* $(1/2, 1/3, 1/6)$*, come in a). Ponendo* $\boldsymbol{p}' = (1/2, 1/2)$ *l'entropia di b) è scritta come una combinazione lineare dell'entropia di eventi e sottoeventi come* $S[\boldsymbol{p}'] + \frac{1}{2} S[(2/3, 1/3)]$ *e si impone che questa entropia sia uguale a quella di a), ovvero* $S[\boldsymbol{p}]$*.*
One of Shannon's requirement for the entropy function. a) Three events (dots) with probabilities $\boldsymbol{p} = (1/2, 1/3, 1/6)$*, and entropy* $S[\boldsymbol{p}]$*. b) The events in a) highlighted in red are replaced by one event with probability* $1/2$*, which is broken into two subevents with probabilities* $2/3$ *and* $1/3$*. Overall, in b) there are three possible events, which occur with probabilities* $(1/2, 1/3, 1/6)$*, as in a). Setting* $\boldsymbol{p}' = (1/2, 1/2)$ *he entropy of b) is written as a linear combination of the entropy of events and subevents as* $S[\boldsymbol{p}'] + \frac{1}{2} S[(2/3, 1/3)]$*, and it is required to equal that of a), i.e.,* $S[\boldsymbol{p}]$*.*

riferisce all'atto di `radere`, e quindi eliminare, caratteristiche superflue e non necessarie in una soluzione. L'idea del rasoio di Ockham può essere facilmente associata all'entropia. Consideriamo a questo proposito un fenomeno dato da N possibili eventi, che si verificano con una probabilità $\boldsymbol{p}$, che vogliamo determinare. Poiché l'entropia $S[\boldsymbol{p}]$ rappresenta la struttura, o complessità, di $\boldsymbol{p}$, il rasoio di Ockham implica che la distribuzione che è più probabile che sia corretta è quella con l'entropia maggiore

$$\begin{cases} \max_{\boldsymbol{p}} S[\boldsymbol{p}], \\ \text{soggetto a} \\ \sum_i p_i = 1. \end{cases} \tag{2}$$

Nella terza riga di Eq. (2) abbiamo incluso la condizione di normalizzazione per $\boldsymbol{p}$. Ciò costituisce un esempio illustrativo di come il principio del rasoio di Ockham sia implementato in pratica: cerchiamo la soluzione più semplice rimuovendo tutti i presupposti superflui e mantenendo solo quelli strettamente necessari. In questo caso, la condizione di normalizzazione della probabilità è l'unico assunto fondamentale che viene preso in considerazione.
L'equazione (2) è la più semplice formulazione illustrativa di un metodo di inferenza statistica, noto

'shaving off' superfluous and unnecessary features in a solution. The idea of Ockham's razor may be naturally related to the entropy. In this regard, let us consider a phenomenon given by N possible events, which are assumed to occur with a probability $\boldsymbol{p}$, which we want to determine. Because the entropy $S[\boldsymbol{p}]$ represents the amount of structure, or complexity, of $\boldsymbol{p}$, Ockham's razor implies that the distribution which is most likely to be correct is the one with the largest entropy

$$\begin{cases} \max_{\boldsymbol{p}} S[\boldsymbol{p}], \\ \text{subject to} \\ \sum_i p_i = 1. \end{cases} \tag{2}$$

In the third line of Eq. (2), we included the normalization condition for $\boldsymbol{p}$. This constitutes a useful, illustrative example of how the Ockham's razor principle is practically implemented: we seek for the simplest solution by removing all unnecessary assumptions, and keeping only those that are strictly necessary. Here, the normalization condition for the probability is the only fundamental assumption which is taken into account.
Equation (2) is the simplest, illustrative mathematical formulation of a statistical-inference method, known

come metodo di massima entropia (MME). Originariamente introdotto da E. T. Jaynes [3], il MME consiste nel cercare la distribuzione di probabilità meno strutturata—o con la massima entropia—che sia coerente con un insieme di condizioni. In questo esempio, la condizione più semplice appare nella terza riga di Eq. (2) ed è data dalla normalizzazione della distribuzione di probabilità.

In generale, nei MME si possono imporre altre condizioni, a seconda del fenomeno specifico in esame. A questo proposito, si consideri un'osservabile fisica, $\mathscr{O}_i$, che dipende dall'evento i. Ad esempio, se il fenomeno in esame è un dado che viene lanciato ed i etichetta il lato del dado che si trova sulla sua superficie superiore, allora $\mathscr{O}_i$ può indicare il numero scritto sul lato i. Se vengono effettuati T lanci del dado, la media empirica di $\mathscr{O}$ è data da

$$\langle \mathscr{O} \rangle_{\text{exp}} = \frac{1}{T} \sum_{t=1}^{T} \mathscr{O}_{i(t)}, \tag{3}$$

dove $i(t)$ denota il lato del dado ottenuto nel t-esimo lancio. Il MME permette di ricostruire $\boldsymbol{p}$ secondo il seguente principio:

$$\begin{cases} \max_{\boldsymbol{p}} S[\boldsymbol{p}], \\ \text{soggetto a} \\ \langle \mathscr{O} \rangle_{\boldsymbol{p}} = \langle \mathscr{O} \rangle_{\text{exp}}, \\ \sum_i p_i = 1, \end{cases} \tag{4}$$

dove

$$\langle \mathscr{O} \rangle_{\boldsymbol{p}} = \sum_{i=1}^{N} p_i \mathscr{O}_i \tag{5}$$

è la media di $\mathscr{O}$ ottenuta con la distribuzione $\boldsymbol{p}$.

Nell'Eq. (4) abbiamo supposto che, assieme alla condizione di normalizzazione, la media di $\mathscr{O}$ costituisca l'osservabile minimale da includere nel modello. Per imporlo, nella terza riga di Eq. (4) richiediamo che la media di $\mathscr{O}$ ottenuta dal modello, $\langle \mathscr{O} \rangle_{\boldsymbol{p}}$, corrisponda a quella sperimentale, $\langle \mathscr{O} \rangle_{\text{exp}}$.
Procedendo come sopra, la formulazione ME (4) può essere facilmente generalizzata a più osservabili, $\mathscr{O}, \mathscr{Q}, \cdots$, imponendo che le medie del modello corrispondano alle rispettive stime sperimentali.

È importante sottolineare che, se il numero T di campioni sperimentali fosse abbastanza grande, si potrebbe campionare direttamente p_i dai dati empirici, senza ricorrere al MME. Tuttavia, in una varietà di istanze sperimentali, il numero di campioni è suf-

as the maximum-entropy method (MEM). Originally introduced by E. T. Jaynes [3], the MEM consists in seeking the least structured—or maximum-entropy—probability distribution which is consistent with a set of conditions. In this example, the simplest condition appears in the third line of Eq. (2), and is given by the normalization of the probability distribution.

More generally, other conditions may be imposed according to the specific phenomenon under consideration. In this regard, consider a physical observable, or 'feature', $\mathscr{O}_i$, which depends on the event i. For instance, if the phenomenon under consideration is a die which is rolled and i labels the side of the die which lies on its upper surface, then $\mathscr{O}_i$ may denote the number marked on side i. If T experimental trials are made, then the experimental average of $\mathscr{O}$ is given by

$$\langle \mathscr{O} \rangle_{\text{exp}} = \frac{1}{T} \sum_{t=1}^{T} \mathscr{O}_{i(t)}, \tag{3}$$

where $i(t)$ denotes the side of the die obtained in the t-th trial. The MEM allows one to reconstruct $\boldsymbol{p}$ according to the following principle:

$$\begin{cases} \max_{\boldsymbol{p}} S[\boldsymbol{p}], \\ \text{subject to} \\ \langle \mathscr{O} \rangle_{\boldsymbol{p}} = \langle \mathscr{O} \rangle_{\text{exp}}, \\ \sum_i p_i = 1, \end{cases} \tag{4}$$

where

$$\langle \mathscr{O} \rangle_{\boldsymbol{p}} = \sum_{i=1}^{N} p_i \mathscr{O}_i \tag{5}$$

is the average of $\mathscr{O}$ obtained with the distribution $\boldsymbol{p}$. In Eq. (4), we assumed that, together with the normalization condition, the average of $\mathscr{O}$ constitutes the minimal feature that needs to be included in the model. To enforce this, in the third line of Eq. (4) we impose that the average of $\mathscr{O}$ obtained from the model, $\langle \mathscr{O} \rangle_{\boldsymbol{p}}$, matches the experimental one, $\langle \mathscr{O} \rangle_{\text{exp}}$. Proceeding along the same lines as above, the ME formulation (4) can be easily generalized to multiple observables, $\mathscr{O}, \mathscr{Q}, \cdots$, whose model averages are required to match their respective experimental estimates.

It is important to point out that, if the number T of experimental trials were large enough, one would be able to directly sample p_i from the empirical data, with no need to resort to the MEM. However, in a variety of experimental instances, the number of sam-

ficientemente grande per stimare solo le medie di un numero finito di osservabili [4, 5], ad esempio, $\langle \mathscr{O} \rangle_{\rm exp}$, non l'intera distribuzione $\boldsymbol{p}$. Di conseguenza, il MME costituisce un metodo pratico per ricostruire, in modo approssimato, la distribuzione di probabilità in presenza di una quantità limitata di dati, facendo leva sull'ipotesi che le informazioni rilevanti siano contenute in una quantità fondamentale, come ad esempio la media di un'osservabile $\mathscr{O}$.
Dal punto di vista matematico, il MME (4) è un problema di ottimizzazione vincolata rispetto alle variabili $p_1, \cdots, p_N$, risolvibile con il metodo dei moltiplicatori di Lagrange. La funzione Lagrange è

$$\mathscr{L}[\boldsymbol{p}] = S[\boldsymbol{p}] - \lambda \left(\langle \mathscr{O} \rangle_{\boldsymbol{p}} - \langle \mathscr{O} \rangle_{\rm exp} \right) - \mu \left(\sum_i p_i - 1 \right), \tag{6}$$

e le condizioni di stazionarietà di $\mathscr{L}$ rispetto a p_i, λ e μ danno

$$\begin{aligned} p_i &= e^{-\lambda \mathscr{O}_i - \mu - 1}, & (7) \\ \langle \mathscr{O} \rangle_{\boldsymbol{p}} &= \langle \mathscr{O} \rangle_{\rm exp}, & (8) \\ \sum_i p_i &= 1, & (9) \end{aligned}$$

rispettivamente, dove nella prima riga abbiamo usato l'Eq. (5). Risolvendo le Eq.ni (7) - (9) rispetto a $\boldsymbol{p}$, λ e μ e sostituendo la soluzione in Eq. (7), si ottiene la distribuzione di probabilità ME $\boldsymbol{p}$.

L'equazione (7) mostra che la distribuzione di probabilità ME dipende esplicitamente dalla scelta della caratteristica $\mathscr{O}$. Inoltre, osserviamo che la forma esponenziale di $\boldsymbol{p}$ ricorda la distribuzione di probabilità di Boltzmann in meccanica statistica [6]. Infatti, uno degli esempi tipici di utilizzo del MME consiste nella derivazione della distribuzione di Boltzmann stessa. Infatti, è stato dimostrato [3] che, se l'indice i etichetta uno stato di un sistema fisico ed $\mathscr{E}_i$ è la sua energia interna, allora la distribuzione ME coerente con il vincolo che l'energia interna media è uguale ad E è la distribuzione di Boltzmann:

$$p_i = \frac{1}{Z} e^{-\mathscr{E}_i/(k_{\rm B}T)}. \tag{10}$$

Nella relazione qui sopra, $Z = \sum_i e^{-\mathscr{E}_i/(k_{\rm B}T)}$ è la funzione di partizione, $k_{\rm B}$ la costante di Boltzmann, e la temperatura inversa $1/(k_{\rm B}T)$ coincide con il moltiplicatore di Lagrange per il vincolo $\langle \mathscr{E} \rangle_{\boldsymbol{p}} = E$, che mette implicitamente in relazione energia interna e temperatura.

ples is sufficient to estimate only averages of a finite number of features [4, 5], e.g., $\langle \mathscr{O} \rangle_{\rm exp}$, not the entire distribution $\boldsymbol{p}$. As a result, the MEM may be used as a practical tool to approximately reconstruct the probability distribution in the presence of a limited amount of data, by leveraging the hypothesis that the relevant information is encoded into a fundamental, minimal quantity such as the average of a feature $\mathscr{O}$.

From the mathematical standpoint, the MEM (4) is a constrained optimization problem with respect to the variables $p_1, \cdots, p_N$, which can be solved with the method of Lagrange multipliers. The Lagrange function reads

$$\mathscr{L}[\boldsymbol{p}] = S[\boldsymbol{p}] - \lambda \left(\langle \mathscr{O} \rangle_{\boldsymbol{p}} - \langle \mathscr{O} \rangle_{\rm exp} \right) - \mu \left(\sum_i p_i - 1 \right), \tag{6}$$

and the stationarity conditions of $\mathscr{L}$ with respect to p_i, λ and μ yield

$$\begin{aligned} p_i &= e^{-\lambda \mathscr{O}_i - \mu - 1}, & (7) \\ \langle \mathscr{O} \rangle_{\boldsymbol{p}} &= \langle \mathscr{O} \rangle_{\rm exp}, & (8) \\ \sum_i p_i &= 1, & (9) \end{aligned}$$

respectively, where in the first line we used Eq. (5). By solving Eqs. (7)-(9) for $\boldsymbol{p}$, λ and μ, and substituting the solution in Eq. (7), one obtains the ME probability distribution p_i.

Equation (7) shows that the ME probability distribution explicitly depends on the choice of the feature $\mathscr{O}$. In addition, we observe that the exponential shape of $\boldsymbol{p}$ is reminiscent of the Boltzmann's probability distribution in statistical mechanics [6]. In fact, one of the prototypical examples of the use of the MEM consists in the derivation of the Boltzmann's distribution itself. Indeed, it has been shown [3] that, if index i labels a state of a physical systems and $\mathscr{E}_i$ is its internal energy, then the ME distribution consistent with the constraint that the average internal energy is equal to E is the Boltzmann distribution:

$$p_i = \frac{1}{Z} e^{-\mathscr{E}_i/(k_{\rm B}T)}. \tag{10}$$

In the relation above, $Z = \sum_i e^{-\mathscr{E}_i/(k_{\rm B}T)}$ is the partition function, $k_{\rm B}$ is Botlzmann's constant, and the inverse temperature $1/(k_{\rm B}T)$ coincides with the Lagrange multiplier for the constraint $\langle \mathscr{E} \rangle_{\boldsymbol{p}} = E$, which implicitly relates internal energy and temperature.

Avendo applicazioni in molteplici campi, i MME si prestano ad essere utilizzati in particolare nel campo dell'intelligenza artificiale. Infatti, la definizione di un 'agente' intelligente come un'entità che percepisce il suo ambiente ed intraprende azioni in modo da massimizzare la probabilità di raggiungere un obiettivo [7], suggerisce un'analogia diretta con l'inferenza statistica ed i MME. In altre parole, l'agente costruisce un modello della realtà basato su un insieme di input esterni, così come i MME costruiscono il modello minimale (7) basato sui dati sperimentali $\langle \mathscr{O} \rangle_{\text{exp}}$ che vengono inseriti in esso. Successivamente, questo modello può essere utilizzato dall'agente per prevedere il comportamento futuro del sistema in esame e, sfruttando queste previsioni, per prendere una decisione al fine di raggiungere un obiettivo specifico.

Among their applications in multiple fields, a notable use of MEMs concerns the domain of artificial intelligence. In fact, the definition of an intelligent 'agent' as an entity which perceives its environment and takes actions so as to maximize the probability of achieving a goal [7], suggests a direct analogy with statistical inference and MEMs. Namely, the agent builds a model of the reality based on a set of external inputs, in the same way in which MEMs build the minimal model (7) based on the experimental data $\langle \mathscr{O} \rangle_{\text{exp}}$ which is fed into it. Later on, this model may be used by the agent to predict the future behavior of the system under consideration and, by leveraging these predictions, to successfully tailor a decision in order to achieve a specific goal.

Punti deboli e critiche

Basi concettuali

Una critica fondamentale che può essere rivolta ai MME riguarda la loro logica concettuale. In effetti, l'assenza di ipotesi superflue nel principio del rasoio di Ockham viene spesso presentata come un punto di forza del metodo, secondo l'idea che nessun *bias* soggettivo, né alcun ingrediente superfluo, vengano introdotti nel modello, ad eccezione dei i dati sperimentali stessi.
Tuttavia, l'assenza di ipotesi può essere considerata essa stessa un'ipotesi non banale. Questo punto può essere illustrato con il MME per uno stormo di uccelli [8]. In breve, la forza di gravità cui sono soggetti gli uccelli indica che non tutte le direzioni spaziali sono equivalenti per lo stormo: di conseguenza, se non si include esplicitamente il ruolo speciale svolto dalla direzione verticale nelle osservabili del MME, si fa, di fatto, un'ipotesi non banale, che potrebbe influenzare i risultati dell'inferenza.

Scelta delle osservabili

Come abbiamo discusso qui sopra, i MME si basano sull'ipotesi che le informazioni fenomenologiche fondamentali siano incluse nella media di un'osservabile $\mathscr{O}$. Tuttavia, la scelta di questa osservabile è dettata dall'intuizione fisica di colui che studia il fenomeno. Ad esempio, per uno stormo di uccelli che volano coerentemente con direzioni di moto quasi parallele, una possibile scelta è data dalla correlazione e

Issues and criticisms

Conceptual basis

A fundamental criticism which may be adressed to MEMs concerns its conceptual rationale. In fact, the absence of superfluous assumptions in Ockham's razor principle is often presented as a selling point of the method, along with the idea that no subjective bias, nor superfluous ingredients, are introduced in the model except for the data itself.
However, the absence of assumptions may be regarded as a nontrivial assumption itself. This point may be illustrated with the MEM for a flock of birds [8]. In short, the force of gravity to which birds are subject indicates that not all spatial directions are equivalent for the flock: as a result, if one does not explicitly include the special role played by the vertical direction in the MEM features, one may be ultimately making a nontrivial assumption, which could bias the results of the ME analysis.

Choice of features

As we discussed above, the MEM is based on the hypothesis that the key phenomenological information is included in the average of a feature $\mathscr{O}$. However, the choice of this feature is dictated by one's physical intuition on the phenomenon under consideration. For instance, for a flock of birds which fly coherently with nearly parallel directions of motion, a possible choice is given by the velocities' correlation and polarization [8]. For the study of collective behavior in

polarizzazione delle velocità [8]. Per lo studio del comportamento collettivo in reti di neuroni, si possono considerare come osservabili i *firing rates* e la funzione di correlazione per i gli *spikes* [9]. Tuttavia, è importante sottolineare che queste ipotesi non sono uniche e sono dettate non solo dalla prospettiva soggettiva di chi osserva il fenomeno, ma anche dalle caratteristiche fisiche che si vogliono studiare nell'esperimento. Come mostrato in Eq. (7), la distribuzione di probabilità di ME dipende dalla scelta di queste osservabili: ne segue che il risultato del MME deve essere sempre considerato con spirito critico e sottoposto a verifica per valutarne l'affidabilità.
Inoltre, un dato insieme di osservabili minimali per il MME potrebbe non essere sufficiente per descrivere correttamente la fenomenologia. Un semplice esempio illustrativo di questa situazione è costituito dal MME per una popolazione di neuroni. Se solo il *firing rate* di ogni neurone venisse incluso nel metodo come osservabile, la distribuzione ME risultante sarebbe data dal prodotto di distribuzioni di *spikes* indipendenti dei neuroni [9]. Ne segue che tale distribuzione ME non potrebbe descrivere alcun comportamento collettivo della rete. Per descrivere questo comportamento collettivo è necessario includere osservabili aggiuntive, come la correlazione a coppie tra gli *spikes* [9]. Tuttavia va ricordato che, anche in presenza di questa osservabile, i risultati del MME potrebbero essere ulteriormente alterati —sia quantitativamente che qualitativamente—se vi si includessero altre osservabili. In altre parole, i risultati ME andrebbero considerati esatti solo se si considerasse un numero infinito di osservabili indipendenti.

Interpretazione

Osservando l'Eq. (7), si può essere tentati di sfruttare l'equivalenza con una distribuzione di Boltzmann ed interpretare $\mathscr{O}_i$ come l'energia del sistema associata allo stato i. Ad esempio, se gli stati del sistema formassero un insieme continuo e fossero etichettati da una variabile reale x, allora si sarebbe tentati di interpretare $\mathscr{O}(x)$ come l'Hamiltoniana del sistema in esame e di mettere in relazione $d\mathscr{O}/dx$ con una forza. Procedendo lungo questa linea, potremmo essere indotti ad affermare che la dinamica temporale del sistema sia data da un moto Browniano nel potenziale $\mathscr{O}(x)$.
Tuttavia, le interpretazioni qui sopra non sarebbero necessariamente corrette [10]. Ricordiamo infatti che

networks of neurons, one may consider as a feature the correlation function for the neural spikes and the firing rates [9]. However, it is important to point out that these hypotheses are not unique, and they are dictated not only by one's subjective perspective on the phenomenon, but also by the specific physical feature that one aims to characterize in the experiment. As shown in Eq. (7), the ME probability distribution depends on the choice of such features: As a resut, the outcome of the MEM must always be regarded with a critical spirit, and subjected to tests in order to assess its reliability.
In addition, a given set of chosen minimal features included in the MEM may not suffice to correctly describe the phenomenology. A simple, illustrative example of this situation consists in the MEM for a population of spiking neurons. If only the spiking frequencies of each neuron are included as features, then the resulting ME distribution is given by the product of independent spiking distributions of the neurons in the network, and it would fail to describe any collective behavior of the neural network as a whole [9]. In order to describe this collective behavior, additional features need to be included, such as the pairwise correlation between spikes [9]. However, it should be reminded that, even in the presence of this feature, the ME results may be further altered—both quantitatively and qualitatively—if additional features were included, and such results should be considered to be exact only if an infinitely large number of independent features were taken into account.

Interpretation

When we look at Eq. (7), it is tempting to leverage the equivalence with a Boltzmann distribution, and interpret $\mathscr{O}_i$ as the energy of the system associated with state i. For instance, if the states of the system formed a continuum set and were labeled by a real variable x, then one would be tempted to interpret $\mathscr{O}(x)$ as the Hamiltonian of the system under consideration, and to relate $d\mathscr{O}/dx$ to a force. Proceeding along the same lines, one would be induced to state that the system dynamics is given by a Browninan motion in the potential $\mathscr{O}(x)$.
However, the interpretations above need not be correct [10]. In fact, we recall that the function $\mathscr{O}(x)$ which appears in the exponential of the ME distribu-

la funzione $\mathscr{O}(x)$ che appare nell'esponenziale della distribuzione ME (7) è semplicemente il risultato di una costruzione matematica, ovvero l'ottimizzazione vincolata (4) e che essa dipende dalla scelta arbitraria delle osservabili del MME.
Di conseguenza, non è garantito che $\mathscr{O}(x)$ abbia alcun significato fisico, né che essa sia connessa ad un'Hamiltoniana o ad una forza. Inoltre, poiché ci sono infiniti processi dinamici che danno origine alla stessa distribuzione stazionaria [11], come ad esempio (7), la dinamica Browniana nel potenziale $\mathscr{O}(x)$ non rappresenta necessariamente la dinamica fisica del sistema.
Un esempio illustrativo del problema qui sopra è dato dai modelli ME per le reti neurali [9]. L'attività della cellula i è rappresentata da una variabile binaria $\sigma_i = \pm 1$, dove '$+1$' significa che la cellula emette uno *spike* e '-1' che resta silenziosa. Un evento è caratterizzato dalla configurazione $\sigma = (\sigma_1, \sigma_2, \cdots)$ della rete e le osservabili per il MME sono i 'firing rates' delle cellule $\sigma_1, \sigma_2, \cdots$ ed i prodotti $\sigma_1\sigma_2, \sigma_1\sigma_3, \cdots$ su tutte le coppie di cellule. Procedendo sulla falsariga dell'Eq. (4), la distribuzione ME è data da

$$p_\sigma \propto \exp\left(-\sum_i h_i\sigma_i - \sum_{i<j} J_{ij}\sigma_i\sigma_j\right), \qquad (11)$$

dove h_i e J_{ij} sono i moltiplicatori di Lagrange per i vincoli.
L'equazione (11) presenta una forte analogia con il modello di Ising in meccanica statistica [12], dove J_{ij} rappresenta il legame fisico, o interazione, tra gli spin i e j. A causa di questa somiglianza, J_{ij} può essere erroneamente interpretato come un'interazione fisica effettiva tra i neuroni i e j. Tuttavia, questa quantità è semplicemente un moltiplicatore di Lagrange nell'ottimizzazione vincolata e non è garantito che rappresenti—quantitativamente né qualitativamente—un'interazione o una connessione fisica tra cellule, come ad esempio una sinapsi.

Errore sperimentale

Un ulteriore punto delicato dei MME riguarda la presenza di incertezze nei dati sperimentali che vengono inseriti nel modello. Dato che le medie sperimentali $\langle\rangle_{\text{exp}}$ risultano da misure, esse possono essere affette da diversi errori, ad esempio strumentali, procedurali, ambientali, umani ed altri. Ad esempio, se il numero di campioni empirici T è abbastanza piccolo, la media $\langle\mathscr{O}\rangle_{\text{exp}}$ in Eq. (3) sarà affetta da un'incertezza

tion (7) is merely the result of a mathematical construction, i.e., the constrained optimization (4), and that it depends on the arbitrary choice of the ME features. As a result, $\mathscr{O}(x)$ is not guaranteed to bear any physical meaning, nor to be related to a Hamiltonian nor a physical force. On top of this, because there are infinitely many dynamical processes that give rise to the same stationary distribution [11] such as (7), the Brownian dynamics in the potential $\mathscr{O}(x)$ need not to represent the actual physical dynamics of the system.
An illustrative example of the issue above is given by ME models for neural networks [9]. The activity of cell i is represented by a binary variable $\sigma_i = \pm 1$, where $+1$ stands for spiking and -1 for being silent. An event is characterized by the configuration $\sigma = (\sigma_1, \sigma_2, \cdots)$ of the network, and the features for the ME construction are the 'spiking rate' of each cell $\sigma_1, \sigma_2, \cdots$, and the products $\sigma_1\sigma_2, \sigma_1\sigma_3, \cdots$ across all cell pairs. Proceeding along the lines of Eq. (4), the ME distribution reads

$$p_\sigma \propto \exp\left(-\sum_i h_i\sigma_i - \sum_{i<j} J_{ij}\sigma_i\sigma_j\right), \qquad (11)$$

where h_i and J_{ij} are the Lagrange multipliers for the constraints. Equation (11) bears a strong similarity to the Ising model in statistical mechanics [12], where J_{ij} represents the physical bond, or interaction, between spins i and j. Because of this similarity, J_{ij} may be misinterpreted as an actual, physical interaction between neurons i and j. Given that this quantity is merely a Lagrange multiplier in the constrained optimization, it is not guaranteed to represent—quantitatively nor qualitatively—physical interactions or connections between neural cells, such as a synapses.

Experimental uncertainties

A further issue with the MEM concerns the presence of uncertainties in the experimental data which is fed into the model. Given that the experimental averages $\langle\rangle_{\text{exp}}$ result from measurements, they may be affected by different sources of errors, e.g., instrumental, procedural, environmental, human, and others. For instance, if the number of empirical samples T is small enough, then the average $\langle\mathscr{O}\rangle_{\text{exp}}$ in Eq. (3) will be

significativa, data dall'errore standard della media. Di conseguenza, imporre che il vincolo di uguaglianza $\langle \mathscr{O} \rangle_{\boldsymbol{p}} = \langle \mathscr{O} \rangle_{\text{exp}}$ in Eq. (4) sia soddisfatto esattamente costituirebbe un criterio troppo rigido e potrebbe produrre risultati errati [13, 14].
Un esempio illustrativo di questo problema viene dai modelli di ME per la modellizzazione del linguaggio [14]. In questo caso, gli eventi sono dati dall'osservazione di coppie (w, w') di parole consecutive, w e w', in un testo. Date due parole, ad esempio, 'saint' e' George', consideriamo due osservabili $\mathscr{O}$ e $\mathscr{Q}$: La frequenza con cui si verifica 'George' nel testo

$$\mathscr{O}_{w,w'} = \mathbb{I}(w' = \text{George}), \tag{12}$$

e la frequenza della coppia 'saint George':

$$\mathscr{Q}_{w,w'} = \mathbb{I}(w = \text{saint}, w' = \text{George}), \tag{13}$$

dove la funzione indicatrice $\mathbb{I}$ è uguale ad uno se tutte le condizioni nel suo argomento sono soddisfatte mentre vale zero in caso contrario. Se si applica l'analisi ME ad un breve testo in cui la parola 'George' ricorre solo dopo 'saint' e si impongono questi vincoli nella loro forma di uguaglianza sulla falsariga di Eq. (4), è semplice dimostrare che la distribuzione ME soddisfa

$$p_{w,\,\text{George}} = 0 \text{ if } w \neq \text{saint}. \tag{14}$$

Questi eventi a frequenza nulla possono causare instabilità numeriche nella stima ME [13]. È ancora più importante ricordare che tali eventi possono provocare scarse prestazioni del modello ME: per esempio, se la distribuzione di ME (14) fosse usata per un riconoscimento di testo, allora qualsiasi evento (w, George) in cui w è diverso da 'saint' non sarebbe riconosciuto come una coppia di parole.
Una soluzione per superare il problema qui sopra consiste nell'allentare i vincoli di uguaglianza nel MME [14]. Ad esempio, supponiamo che i dati non siano sufficientemente accurati da fornire un valore per la media sperimentale, ma che essi possano stimare solo un intervallo di confidenza dato da un limite superiore e inferiore $\mathscr{O}_+$ e $\mathscr{O}_-$, rispettivamente:

$$\mathscr{O}_- \leq \langle \mathscr{O} \rangle_{\text{exp}} \leq \mathscr{O}_+. \tag{15}$$

A questo punto si può cercare la distribuzione meno strutturata $\boldsymbol{p}$, che è coerente con questa informazione sperimentale, riformulando il metodo ME (4) come

affected by a significant uncertainty, related to its standard error of the mean. As a result, a full satisfaction of the equality constraint $\langle \mathscr{O} \rangle_{\boldsymbol{p}} = \langle \mathscr{O} \rangle_{\text{exp}}$ in Eq. (4) would be too strict of a criterion, and it may produce misleading results [13, 14].
An illustrative example of this issue comes from ME models for language modeling [14]. In this case, the events are given by the observation of pairs (w, w') of consecutive words, w and w', in a corpus of text. Given two words, e.g., 'saint' and 'George', we consider two features $\mathscr{O}$ and $\mathscr{Q}$: The frequency with which 'George' occurs in the text

$$\mathscr{O}_{w,w'} = \mathbb{I}(w' = \text{George}), \tag{12}$$

and the frequency of the bigram 'saint George', i.e.,

$$\mathscr{Q}_{w,w'} = \mathbb{I}(w = \text{saint}, w' = \text{George}), \tag{13}$$

where the indicator function $\mathbb{I}$ is one if all conditions in its argument are satisfied, and zero otherwise. If one applies the ME analysis to a short corpus of text where the word 'George' occurs only after 'saint', and imposes these constraints in their equality form along the lines of Eq. (4), it is straightforward to show that the ME distribution satisfies

$$p_{w,\,\text{George}} = 0 \text{ if } w \neq \text{saint}. \tag{14}$$

These zero-frequency events in the ME model may cause numerical instabilities in ME estimation [13]. More importantly, such events may result in poor performance of the ME model: For instance, if the ME distribution (14) were used for text recognition, then any word pair (w, George) in which w is different from 'saint' would not be recognized as a bigram.
A solution to overcome the issue above consists in relaxing the equality constraints in the MEM [14]. For instance, let us suppose that the data is not accurate enough to provide a value for the experimental average, but can only give a confidence interval given by an upper and lower bound $\mathscr{O}_+$ and $\mathscr{O}_-$, respectively:

$$\mathscr{O}_- \leq \langle \mathscr{O} \rangle_{\text{exp}} \leq \mathscr{O}_+. \tag{15}$$

Then one may seek for the least-structured distribution $\boldsymbol{p}$, which is consistent with this experimental information, by reformulating the ME method (4) as

segue:

$$\begin{cases} \max_{\boldsymbol{p}} S[\boldsymbol{p}], \\ \text{soggetto a} \\ \mathscr{O}_- \leq \langle \mathscr{O} \rangle_{\text{exp}} \leq \mathscr{O}_+, \\ \sum_i p_i = 1. \end{cases} \tag{16}$$

Dal punto di vista matematico, Eq. (16) è un problema di massimizzazione con vincoli sia di uguaglianza che di disuguaglianza, che può essere risolto con metodi matematici noti, introdotti da W. Karush, H. W. Kuhn ed A. W. Tucker (KKT) a metà del ventesimo secolo [15, 16]. L'approccio KKT ha una somiglianza con il metodo Lagrange per i vincoli di uguaglianza, Eq. (6): ad ogni vincolo di disuguaglianza è associato un moltiplicatore KKT non negativo. In breve, se il moltiplicatore svanisce, allora il massimo si trova all'interno di una regione nello spazio delle probabilità $\boldsymbol{p}$, dove i vincoli di disuguaglianza sono soddisfatti: di conseguenza, il vincolo di disuguaglianza è irrilevante per la massimizzazione. D'altra parte, se il moltiplicatore è positivo, allora il massimo si trova sul bordo della regione in cui il vincolo è soddisfatto e la presenza del vincolo di disuguaglianza influenza il valore massimo dell'entropia nell'ottimizzazione.

Discussione

Data la loro generalità, versatilità computazionale e semplicità, negli ultimi anni i metodi di massima entropia (MME) sono stati applicati a una grande varietà di fenomeni. Spaziando dal comportamento animale collettivo, alle sequenze in famiglie di proteine, alle parole in un dizionario o testo, la sorprendente facilità con cui i MME possono essere applicati si è tradotta in un gran numero di modelli, in base ai quali si è cercato di capire i meccanismi che sono alla base dei fenomeni in esame. Dopo aver presentato una breve introduzione ai MME, ne abbiamo fornito un'analisi critica, in modo da guidare il lettore attraverso i potenziali svantaggi e punti deboli di questi metodi di inferenza. Sebbene questo elenco di potenziali problemi non sia esaustivo, il nostro obbiettivo è stato di fornire al lettore un'idea generale di quale tipo di critiche e domande dovrebbero e potrebbero essere indirizzate alle analisi ME e come interpretare correttamente le loro conclusioni sul fenomeno in esame.

follows:

$$\begin{cases} \max_{\boldsymbol{p}} S[\boldsymbol{p}], \\ \text{subject to} \\ \mathscr{O}_- \leq \langle \mathscr{O} \rangle_{\text{exp}} \leq \mathscr{O}_+, \\ \sum_i p_i = 1. \end{cases} \tag{16}$$

From the mathematical standpoint, Eq. (16) is a maximization problem with both equality and inequality constraints, which can be solved with known mathematical methods introduced by W. Karush, H. W. Kuhn and A. W. Tucker (KKT) in the mid-twentieth century [15, 16]. The method bears a similarity to the Lagrange method for equality constraints, Eq. (6): to each inequality constraint is associated a non-negative KKT multiplier. In short, if the multiplier vanishes, then the optimum lies within a region in the space of probabilities $\boldsymbol{p}$, where the inequality constraints is satisfied: as a result, the inequality constraint is irrelevant for the optimum. On the other hand, if the multiplier is positive, then the optimum is located at the boundary of the region where the constraint is satisfied, and the presence of the inequality constraint lowers the optimal value of the objective function.

Discussion

Given their generality, computational versatility and straightforwardness, in recent years maximum-entropy methods (MEMs) have been widely applied to a large variety of phenomena. Spanning from collective animal behavior, to sequences in families of proteins, to words in a dictionary or corpus of text and others, the striking easiness which MEMs can be applied resulted in a plethora of proposed models and interpretations on the underlying mechanisms of the phenomena under consideration. After presenting a short introduction to MEMs, we provided a critical assessment of MEMs, so as to guide the general reader through the potential drawbacks and weak spots of this inference method. While this list of potential issues is not meant to be exhaustive, we aimed at providing the reader with a general flavor of what kind of criticisms and questions should and could be adressed to ME analyses, and how to correctly interpret their results and claims on the phenomenon under consideration.

[1] C.E. Shannon: *A mathematical theory of communication*, Bell System Tech. J. 27 (1987) 379.

[2] E. Sober: *Ockham's Razors: A User's Manual*, Cambridge Univ. Press, Cambridge (UK) (2015).

[3] E.T. Jaynes: *Information theory and statistical mechanics*, Physical Review 106 (1957) 620.

[4] S. Cocco, R. Monasson, M. Weigt: *From Principal Component to Direct Coupling Analysis of Coevolution in Proteins: Low-Eigenvalue Modes are Needed for Structure Prediction*, PLoS Comput. Biol. 8 (2013) e1003176.

[5] M. Weigt, et al., *Identification of direct residue contacts in protein–protein interaction by message passing*, P. Natl. Acad. Sci. USA 106 (2009) 67.

[6] L.D. Landau, E.M. Lifshitz: *Statistical Physics: Course of Theoretical Physics*, Pergamon Press, London (1980).

[7] D. Poole, A. Mackworth, R. Goebel: *Computational Intelligence: a Logical Approach*, Oxford University Press, Oxford (1998).

[8] W. Bialek, et al.: *Flocking is a typical example of emergent collective behavior, where interactions between individuals produce collective patterns on the large scale.*, P. Natl. Acad. Sci. USA 109 (2012) 4786.

[9] E. Schneidman, et al.: *Weak pairwise correlations imply strongly correlated network states in a neural population*, Nature 440 (2006) 1007.

[10] M. Castellana, W. Bialek, A. Cavagna, I. Giardina: *Entropic effects in a nonequilibrium system: Flocks of birds*, Phys. Rev. E 93 (2016), 052416.

[11] P.C. Hohenberg, B.I. Halperin: *Theory of dynamic critical phenomena*, Rev. Mod. Phys. 49 (1977) 435.

[12] K. Huang: *Statistical Mechanics*, Wiley Press, New York (1987).

[13] J. Kazama, J. Tsujii: *Maximum entropy models with inequality constraints: A case study on text categorization*, Mach. Learn. 60 (2005) 159.

[14] S.F. Chen, R. Rosenfeld: *A survey of smoothing techniques for ME models*, IEEE T. Speech Audi. P. 8 (2000) 37.

[15] W. Karush: *Minima of functions of several variables with inequalities as side constraints*, Chicago Univ. (Math. Dept.) (1939).

[16] H.W. Kuhn, A.W. Tucker: *Second Berkeley Symposium on Mathematical Statistics and Probability*, Non-Linear Program. University of California Press, Berkeley (1951).

Michele Castellana: è ricercatore presso il Laboratoire Physico-Chimie Curie di Parigi, un'unità di ricerca del centro nazionale francese di ricerca scientifica (CNRS) che fa parte dell'Institut Curie. Si è laureato in fisica teorica presso l'Università La Sapienza di Roma, con una tesi su approcci di teoria di campo alla gravità quantistica. Per il suo dottorato di ricerca, svoltosi tra l'Università La Sapienza e l'Università Paris Sud, è passato alla fisica statistica e si è concentrato sui metodi di gruppo di rinormalizzazione per sistemi disordinati. Si è poi trasferito all'università di Princeton per il post-dottorato, dove ha lavorato su argomenti tra fisica statistica e biologia.

Michele Castellana: is an associate scientist at Laboratoire Physico-Chimie Curie in Paris, a research unit of the French National Centre for Scientific Research (CNRS) which is part of Institut Curie. He graduated in theoretical physics from Sapienza University of Rome, with a thesis on field-theoretical aspects of quantum gravity. For his Ph.D., joint between Sapienza University and University Paris Sud, he switched to statistical physics, and focused on renormalization-group methods for disordered systems. He then moved to Princeton University as a postdoctoral associate, where he worked on topics at the boundary between statistical and biological physics.

I Computer e il Linguaggio Naturale

You shall know a word by the company it keeps.
R. Firth

Valerio Basile *Dipartimento di Informatica*
Università degli Studi di Torino, Italia

Il linguaggio naturale, con le sue strutture e irregolarità, è il principale oggetto di studio della `Linguistica`, ma anche di `Informatica` e `Intelligenza Artificiale`; interessate analizzarlo, comprenderlo, ed elaborarlo con strumenti computazionali. Le discipline della `Linguistica Computazionale` ed `Elaborazione del Linguaggio Naturale` nascono con lo scopo di coniugare conoscenze da vari campi di studio per svelare tramite computer ed algoritmi i segreti della lingua. In questo articolo, viene presentata una panoramica delle tante sfaccettature di queste aree di ricerca, i problemi che esse affrontano, e le soluzioni proposte negli ultimi decenni di letteratura scientifica.

Introduzione

Il linguaggio naturale è tra le principali espressioni dell'intelligenza umana, una modalità di comunicazione che si è evoluta in forme estremamente diverse e complesse, e che ha suscitato la curiosità degli studiosi fin dagli albori della filosofia. Lo chiamiamo `naturale` per distinguerlo dai linguaggi `artificiali`, ad esempio i linguaggi formali come linguaggi di programmazione, la notazione matematica, o musicale. Allo stesso tempo, il linguaggio naturale è qualcosa che tutti conoscono e usano in modo, appunto, naturale, spesso senza rendersi consciamente conto di seguire delle precise regole linguistiche.

In quanto fenomeno strettamente correlato con l'intelligenza, il linguaggio naturale è ampiamente studiato, oltre che da filosofi, linguisti, scienziati cognitivi, ecc., anche da informatici e ingegneri che si occupano di Intelligenza Artificiale. La **Linguistica Computazionale** (Computational Linguistics in inglese) è quella area di studio e di ricerca che coniuga linguistica e informatica, allo scopo di analizzare il linguaggio naturale con metodi informatici, con modelli statistici, regole formali, e programmi per computer.

Linguistica Computazionale è un termine usato a volte come sinonimo di **Elaborazione del Linguaggio Naturale** (Natural Language Processing, spesso abbreviato in NLP), anche se ci sono ampie discussioni sulle differenze tra queste terminologie. Senza voler prendere una posizione netta, una visione ragionevolmente condivisa sostiene che per Linguistica Computazionale si intende lo studio della lingua supportato da metodologie computazionali, mentre con NLP ci si riferisce alla estesa famiglia di tecniche computazionali che trattano la lingua come dato principale.

In questo articolo, cercherò di fornire delle basi per un'infarinatura generale che copra il più possibile della Linguistica Computazionale, o quantomeno dei problemi più studiati tra quelli di cui questa disciplina si occupa. Realisticamente, l'obiettivo di questo articolo è quello di suscitare nei lettori la scintilla di curiosità che li spinga a investigare oltre, a partire dai numerosi link e puntatori sparsi per l'articolo.

Si rende infine necessario tracciare dei confini riguardo il contenuto del presente articolo, data l'enorme mole di letteratura intorno alla Linguistica Computazionale. Due ampie aree di ricerca sono state volutamente escluse dalla trattazione principale. Innanzitutto, il processamento automatico del linguaggio naturale può naturalmente distinguersi nei due filoni della `comprensione` e della `generazione` del linguaggio naturale (Natural Language Understanding e Generation, rispettivamente, in inglese). In questo articolo, si tratterà principalmente del primo, al cui studio è dedicata una comunità internazionale di gran lunga più numerosa. L'altra distinzione riguarda la lingua `parlata` e la lingua `scritta`. Questo articolo tratterà quasi esclusivamente di tecniche sviluppate per l'elaborazione della lingua scritta.

Linguistica e Computer

La comprensione, ma anche la sola modellazione o processamento automatico, di un'espressione in linguaggio naturale da parte di una macchina è un problema molto complicato. Per questo motivo, l'evoluzione della disciplina ha portato alla sua divisione in molteplici sotto-problemi, la cui soluzione è spesso più gestibile con metodi informatici.

Restringendo il campo alla lingua scritta, dal punto di vista dell'elaboratore eletronico una espressione in linguaggio naturale è, in prima battuta, una sequenza (o `stringa`) di caratteri. Già a questo livello, la sua rappresentazione digitale non è scontata, e infatti sono nati e proliferati una serie di standard come `ASCII` e più recentemente `UNICODE`, allo scopo di convertire i caratteri in codici numerici processabili da un computer. In questo articolo non ci soffermiamo su tali standard di rappresentazione. Piuttosto, nel resto di questa sezione, vedremo come il linguaggio naturale viene processato su livelli via via più astratti a partire dalle sequenze di caratteri fino al significato di espressioni complesse.

Caratteri, parole, punteggiatura e frasi

Per un essere umano in grado di leggere, è scontato guardare ad un testo (come quello che compone il presente articolo) come un insieme di parole e frasi. Tuttavia, questa astrazione non è affatto scontata per un computer, che, come abbiamo scritto, vede il suo input come una stringa di caratteri. Il processo di suddividere un testo in frasi, e le frasi in parole, prende il nome di **tokenizzazione**, da *token*, unità elementare di un testo.

Il problema è all'apparenza banale: basta individuare i segni di punteggiatura che dividono le frasi, e successivamente individuare le parole, separate da spazi o da altri segni di punteggiatura. È così semplice? La risposta è si nella maggior parte dei casi, ma questo semplice insieme di regole lascia fuori una quantità non indifferente di eccezioni. Vediamo un esempio:

> "Giovanni è tornato dagli U.S.A. C'era già stato l'anno scorso."

In questo testo di esempio, ci sono due frasi, composte da 5 e 7 parole rispettivamente (più il punto finale). Si vede come per un sistema basato su semplici regole possa essere problematico separare le coppie di parole "c'era" e "l'anno". ancora più problematiche sono le decisione di separare la prima frase dalla seconda, e mantenere unita l'abbreviazione "U.S.A.": come decidere se i punti fanno parte di un'abbreviazione o sono usati come comuni segni di punteggiatura?

I moderni sistemi di tokenizzazione utilizzano per la maggior parte complessi sistemi di regole, differenti per ogni lingua. Sono stati però proposti anche tokenizzatori più sofisticati, basati su tecniche di apprendimento automatico (più avanti si parlerà più in dettaglio di tali tecniche). Un sistema di tokenizzazione di questo tipo è `Elephant` [2], che impara le regole automaticamente da un *corpus* corretto a mano, indipendentemente dalla lingua. Il nome del sistema fa riferimento alla metafora dell' "elefante nel salotto", alludendo a come la tokenizzazione sia un problema ancora aperto, anche se dato per scontato e risolto in molti contesti.

Lemmi, forme, inflessioni

Quando apriamo un dizionario per cercare un termine, troviamo le `forme base` delle parole di una certa lingua, o cosiddetti `lemmi`. In italiano, ad esempio, la forma base di un verbo è l'infinito, la forma base di un sostantivo è il maschile singolare, e così via. In linguistica, questa distinzione è nota come la differenza tra lemmi e `forme` delle parole. Ad ogni lemma, ad esempio un verbo come "andare" sono associate in genere molte forme differenti, che rispecchiano *feature* come tempi e modi verbali, numero (singolare o plurale), genere (maschile o femminile), ed altre. A seconda della lingua, le `inflessioni morfologiche` dei lemmi sono più o meno numerose e più o meno regolari. Alcune lingue hanno una morfologia più semplice e regolare, come l'inglese[1], altre, come le lingue romanze, più complesse. Tra le lingue considerate più ricche morfologicamente si annoverano usualmente turco e finlandese.

La **morfologia lessicale** (lo studio della forma delle parole) ha sempre interessato i linguisti computazionali, per via della sua natura fatta di regole ed eccezioni. Il *task* principale rispetto alla morfologia in una tipica *pipeline* NLP è quello della **lemmatizzazione**: data in input una parola (quindi una forma in generale flessa), ricavare il corrispondente lemma. Questo task è importante non solo ai fini dell'analisi morfologica in sé, ma anche perché è spesso strumentale ad analisi successive, come la ricerca di parole nel testo da analizzare all'interno di risorse lessicali.

Un approccio semplice ma efficace alla lemmatizzazione consiste nell'implementazione di **formari** computazionali, tabelle che associano ad ogni lemma di una lingua tutte le sue possibili forme flesse. Questo approccio ha il vantaggio di non richiedere computazioni complesse, al di là del mantenimento in memoria della tabella delle forme. Inoltre, in questo modo è possibile gestire le flessioni morfologiche irregolari (es. "andare" → "vado") direttamente, senza bisogno di aggiungere regole particolari. D'altro canto, questo semplice approccio ha un problema di `scalabilità`, poiché diviene difficile considerare termini altamente specifici (es. in ambito bio-

[1] Nella mia tesi di dottorato, ho mostrato come i quattro suffissi *-s, -ed, -ing, -en* coprano la maggior parte della morfologia lessicale della lingua inglese [3]

Forma	Lemma	Feature
gattini	gattino	NOUN-M:p
andarono	andare	VER:ind+past+3+p
fastidiosetto	fastidioso	ADJ:dim+m+s

Tabella 1: *Esempio del contenuto del formario Morph-it! [4].*

medico) o neologismi. Per l'italiano, un formario liberamente utilizzabile è `Morph-it!`[2], contenente 504906 forme flesse per 34968 lemmi [4]. Un esempio del contenuto di `Morph-it!` è riportato in Tabella 1, dove nella terza colonna sono visibili le *feature* morfologiche nominate in precedenza, come parte del discorso, genere, persona, e così via.

Tra i metodi computazionali più studiati in ricerca per l'analisi morfologica troviamo quelli basati su formalismi chiamati **automi a stati finiti**, e più precisamente su **trasduttori**. Questo tipo di formalismi descrive una serie di `stati` e regole di `transizione` tra essi. L'automa prende in *input* sequenze di valori da un vocabolario predefinito (ad esempio lettere, parole, o morfemi) e, a seconda del loro contenuto, vengono attivate diverse regole, formando un percorso all'interno del grafo diretto che costituisce l'automa. Gli automi a stati finiti sono fondamentali, tra le altre cose, per il riconoscimento della correttezza sintattica dei linguaggi formali, come i linguaggi di programmazione. I trasduttori, in particolare, hanno la caratteristica di emettere simboli all'attivazione di ogni regola di transizione.

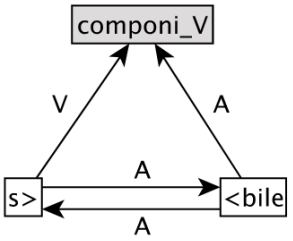


Figura 1: *Esempio di analisi morfologica usando trasduttori a stati finiti. Immagine presa dall'articolo di Tamburini e Melandri [5].*

Un sistema che implementa un modello basato su trasduttori a stati finiti per creare un analizzatore morfologico per l'italiano è `AnIta` [5]. La Figura 1 mostra un grafo che descrive il fram-

[2] `https://docs.sslmit.unibo.it/doku.php?id=resources:morph-it`

mento del sistema capace di riconoscere le parole "comporre", "scomporre", "componibile", e "scomponibile", partendo dalla forma base verbale evidenziata. Le etichette sugli archi in figura indicano il riconoscimento di un verbo (V, come nel caso di "'scomporre') o di un aggettivo (A).

Infine, sono stati proposti approcci alla lemmatizzazione basati su metodi di apprendimento automatico supervisionato. Un esempio tra questi è `morfette` [6], un sistema basato su un classificatore *maximum entropy*, che apprende da una collezione di parole morfologicamente analizzate regole di analisimi morfologica codificate come una sequenza di `correzioni`, come aggiunte, cancellazioni, e sostituzioni di lettere.

Parti del discorso e sintassi

Per molti, il primo approccio con la linguistica, alle scuole elementari, prende la forma dell'analisi grammaticale. Una parte di questo processo è costituito da una versione semplificata dell'analisi morfologica vista nella sezione precedente. L'altro *task* complementare è l'individuazione delle `parti del discorso` (*part of speech*, in inglese). Il *part-of-speech tagging* è il *task* che ha come *input* una sequenza di parole, debitamente separate tra loro (**tokenizzazione**) ed eventualmente lemmatizzate e corredate da informazioni morfologiche, e produce una etichetta (**tag**) per ogni parola che ne identifica la parte del discorso, come ad esempio, verbo, nome, avverbio, articolo determinativo, ecc. Anche questo *task* è importante di per sé, ma anche propedeutico per elaborazioni successive.

Innanzitutto, per etichettare le parole con parti del discorso, bisogna stabilire quali sono le possibilità. Diversi *tagset* sono stati proposti in letteratura, più o meno dipendenti da una certa lingua, ad esempio dalle `Università della Pennsylvania` e di `Standford` per l'inglese. Molte estensioni sono nate per catturare fenomeni specifici di alcune lingue o famiglie di lingue. Recentemente, l'iniziativa `Universal Dependencies`[3] ha lavorato per integrare ed armonizzare i diversi standard, e proporre uno standard, appunto, universale, compreso un tagset di parti del discorso. L'ultima versione (2.5) delle specifiche del il consorzio UD prevedono un *tagset* formato da 17 categorie, di cui 8 sono `classi chiuse`, ovvero contenenti un numero finito di parole per ogni lingua, 6 sono `classi aperte`, e 3 coprono le restanti categorie di *token*. Un esempio di classe chiusa di parole è quella delle preposizioni, un insieme fisso in ogni determinata lingua. Il *tagset* `UD` è raffigurato in Tabella 2. In aggiunta ai 17 *tag*, lo standard `UD` prevede una serie di *feature* per codificare le caratteristiche grammaticali e lessicali aggiuntive, come quelle viste nella sezione sulla morfologia computazionale.

Dato un *tagset* e definito il *task*, come riusciamo tramite metodi computazionali ad assegnare le corrette parti del discorso ad ogni parole di una frase o di un testo? Innanzitutto, notiamo una differenza con il *task* di analisi morfologica che abbiamo visto nella sezione precedente. Mentre per quello può essere sufficiente analizzare una parola alla volta, in questo caso il `contesto` gioca un ruolo fondamentale. L'ambiguità delle parti del discorso è infatti un fattore notevole da prendere in considerazione. Vediamo una frase di esempio:

> "Domani volo a Milano, ma il volo è in ritardo."

In questo esempio, "volo" è un verbo alla prima persona singolare alla sua prima occorrenza, e un sostantivo la seconda. Tuttavia, leggendo la frase non abbiamo dubbi riguardo la sua interpretazione, poiché il contesto rende evidente la giusta attribuzione delle parti del discorso. Alcuni casi, rari, sono effettivamente ambigui anche in presenza dell'intera frase, come la seguente frase conosciuta in linguistica:

> "La vecchia porta la sbarra."

Lascio al lettore il divertente compito di individuare tutti possibili significati di questa frase[4]. Anche senza considerare casi estremi come l'esempio citato, rimane il problema di costruire sistemi automatici per il `POS`-*tagging*. I primi approcci computazionali a questo *task* sono stati basati sulla creazione di regole che associano sequenze di parole con determinate caratteristiche grammaticali ai rispettivi *tag* di parte del

[3] `https://universaldependencies.org/`

[4] Una frase in inglese dalle caratteristiche analoghe è *time flies like an arrow*.

Classi aperte		Classi chiuse		Altri	
ADJ	Aggettivo	ADP	Preposizione	PUNCT	Punteggiatura
ADV	Avverbio	AUX	Ausiliare	SYM	Simbolo
INTJ	Interiezione	CCONJ	Congiunzione coordinata	X	altro
NOUN	Sostantivo	DET	Articolo		
PROPN	Nome proprio	NUM	Numero		
VERB	Verbo	PART	Particella		
		PRON	Pronome		
		SCONJ	Congiunzione subordinata		

Tabella 2: *Il tagset di parti del discorso dello standard Universal Dependencies.*

discorso. Tali regole vanno a formalizzare vere e proprie grammatiche di una lingua, e il loro numero può crescere fino a diventare difficilmente gestibile.

Negli anni '70, grazie al lavoro condotto presso la `Brown University`, è stato reso disponibile il primo *corpus* di testi in lingua inglese annotato con parti del discorso. Il processo di annotazione, condotto a mano da volontari, si è protratto per diversi anni, ma è stato di importanza fondamentale non solo per la creazione di programmi `POS`-*tagger* automatici, ma per lo sviluppo dell'intera disciplina della Linguistica Computazionale. A partire dal `Brown Corpus`, infatti, sono stati sviluppati modelli statistici per il *tagging* di sequenze, basati principalmente su `modelli nascosti di Markov` e `Maximum Entropy`, e, in tempi più recenti, reti neurali. Tali metodi imparano le sequenze di `POS`-*tag* più probabili data una sequenza di input di parole con le loro caratteristiche grammaticali, lessicali e morfologiche, *output* dei *task* visti in precedenza.

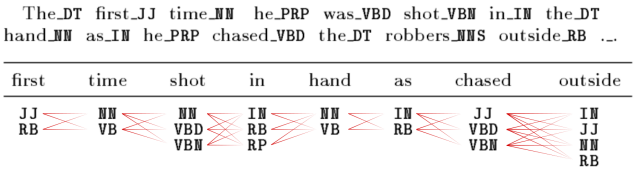


Figura 2: *Ambiguità di parti del discorso e tutte le possibili sequenze di POS-tag. Immagine tratta dall'articolo di Màrquez et al. [9].*

La predizione della corretta sequenza di etichette è interessante dal punto di vista computazionale. Si prenda ad esempio la frase in Figura 2, dove per ciascuna parola sono indicate le possibili parti del discorso. Un sistema come un modello nascosto di Markov mira ad individuare la sequenza più probabile di parti del discorso, quindi, in termini computazionali, il percorso nel grafo sovraimposto alla frase nella figura, che massimizza la probabilità totale. Su ogni arco del grafo si possono infatti immaginare dei numeri che indicano la probabilità di passare da uno stato al successivo, in modo simile al trasduttore a stati finiti visto in precedenza per il *task* di analisi morfologica. Queste probabilità possono essere stimate a partire da un corpus annotato con parti del discorso, come il `Brown Corpus`. I possibili percorsi, tuttavia, rimangono un numero elevato, che cresce `esponenzialmente` al crescere della dimensione dell'*input* (la lunghezza della frase). Già in questa breve frase di esempio le possibili combinazioni di `POS`-*tag* sono $2 \cdot 2 \cdot 3 \cdot 3 \cdot 2 \cdot 2 \cdot 3 \cdot 4 = 1,728$. Sono pertanto stati proposti algoritmi per risolvere questo problema in modo ottimizzato, tra cui probabilmente il più conosciuto è l'algoritmo di `Viterbi` [8], basato sulla tecnica della `programmazione dinamica`.

Nel corso di un paio di decadi, modelli sempre più sofisticati hanno raggiunto valori di accuratezza dal 90% dei primi modelli fino ai valori vicini al 98% dei sistemi più recenti. È comunque interessante notare come una accuratezza del 98% si traduca in circa un errore di assegnamento di parte del discorso ogni 20 parole, una *performance* ancora lontana da quella umana. Allo stato attuale, pertanto, il `POS`-*tagging* è considerato un problema aperto su cui la ricerca rimane attiva e prolifica.

Sintassi del linguaggio naturale

Le parole che compongono un'espressione in linguaggio naturale non appaiono in ordine casuale in una frase. Al contrario, la loro precisa sequenza, e come ogni parola si relaziona alle altre sono fattori fondamentali per comprenderne il significato. La sintassi è l'area della linguistica che studia i rapporti e le relazioni tra le parole, e il modo

in cui questi, insieme, compongono espressioni di senso compiuto.

La letteratura linguistica ha prodotto innumerevoli grammatiche e formalismi per descrivere le regole che permettono alle parole di legarsi tra loro e formare espressioni più lunghe, e non basterebbe un articolo di queste dimensioni neanche a farne una rapida carrellata. Ci limiteremo quindi ad una distinzione a grana grossa tra le due maggiori famiglie di formalismi sintattici del linguaggio naturale.

In generale, l'analisi sintattica di un frammento di linguaggio naturale ha come risultato un `albero`, un caso specifico di grafo in cui ogni nodo ha uno ed un solo nodo `padre`, ad eccezione di un nodo speciale chiamato `radice`, che non ha padre. In nodi situati più "in basso" nell'albero, cioé quei nodi che sono collegati solo ciascuno al proprio nodo padre, sono chiamati `foglie`. Un albero, in informatica, è la maniera naturale di rappresentare informazioni gerarchiche. L'albero che rappresenta la struttura sintattica di una frase è diverso a seconda del tipo di grammatica che si sta utilizzando. Nelle grammatiche a `costituenti` le foglie sono le parole della frase, la radice è la frase intera, e i nodi intermedi sono raggruppamenti di parole, chiamti appunto `costituenti`, che si relazionano tra loro in base alle regole della grammatica usata. Per fare un esempio, guardiamo l'albero a costituenti in Figura 3, dove la radice è il livello più in basso e le foglie sono in alto nei *box* colorati[5]. Qui il formalismo grammaticale utilizzato è la `Combinatory Categorial Grammar`, una grammatica relativamente semplice in termini di regole ma nonostante ciò molto espressiva. In `CCG`, ogni parola ha una `categoria` sintattica che può essere semplice (`N`, `NP`, oppure la categoria speciale `S` che indica l'intera frase), oppure complessa. Le categorie complesse si costruiscono con le categorie semplici più i due operatori \ e / che indicano rispettivamente "questo elemento si compone con un altro elemento a destra" e "a destra". La `CCG` quindi fornisce delle regole su come due costituenti, ognuno con la propria categoria, si combinano per formarne uno nuovo. Ad esempio, NP/N è una categoria complessa che può essere combinata con un N alla propria destra per formare un NP. Seguendo queste regole di composizione, il *parser* (il programma che effettua il *parsing*, ovvero l'analisi sintattica) è capace di ricostruire un albero sintattico completo a partire dalle categorie dei figli.

L'altra famiglia di formalismi sintattici è quella delle grammatiche `a dipendenze`. In questo tipo di grammatiche, Tutti i nodi, compresi radice e foglie, sono parole della frase, e gli archi che le collegano sono etichettati con le loro `dipendenze` ovvero il tipo di relazione che sussiste tra coppie di parole. Un esempio di albero di dipendenze si trova in Figura 4, prodotto dalla versione online del software `UDpipe`, una *pipeline* `NLP` sviluppata all'interno della già citata iniziativa `Universal Dependencies`[6]. Si puo vedere come ad esempio tutta la frase "Tom sta masticando uno stuzzicadenti" `dipenda` dal verbo principale "masticare", il cui soggetto (`nsubj`) è Tom e il cui oggetto (`obj`) è lo stuzzicadenti.

Sia per le grammatiche a costituenti sia per le grammatiche a dipendenze, in letteratura sono stati (e continuano ad essere) proposti un gran numero di algoritmi di `parsing`. Anche per questo *task*, mentre ai principi dello sviluppo della linguistica computazionale sono stati proposti sistemi basati su regole, a partire dagli anni '80 la creazione di corpora annotati con analisi sintattiche (i cosiddetti *treebank*, da *tree*, albero) ha permesso lo sviluppo di metodi basati sulla statistica e l'apprendimento automatico. Senza entrare in eccessivo dettaglio, la maggior parte degli algoritmi più comuni al giorno d'oggi sono di apprendimento automatico supervisionato, basati o sulle `transizioni`, ovvero che cercano di predire la dipendenza tra coppie di parole, o basati su algoritmi che operano sull'intero grafo delle possibili dipendenze tra le parole di una frase.

Le parole e il loro significato

Vedremo ora quei task che riguardano la `semantica`, ovvero lo studio del significato del linguaggio naturale. Il primo livello su cui si può concentrare l'attenzione alla ricerca del si-

[5] Questo esempio è stato estratto dal Parallel Meaning Bank [7], un corpus annotato multilingue liberamente accessibile online presso `https://pmb.let.rug.nl/explorer`.

[6] UDpipe è utilizzabile liberamente, e fornisce analisi linguistica su un gran numero di lingue: http://lindat.mff.cuni.cz/services/udpipe/

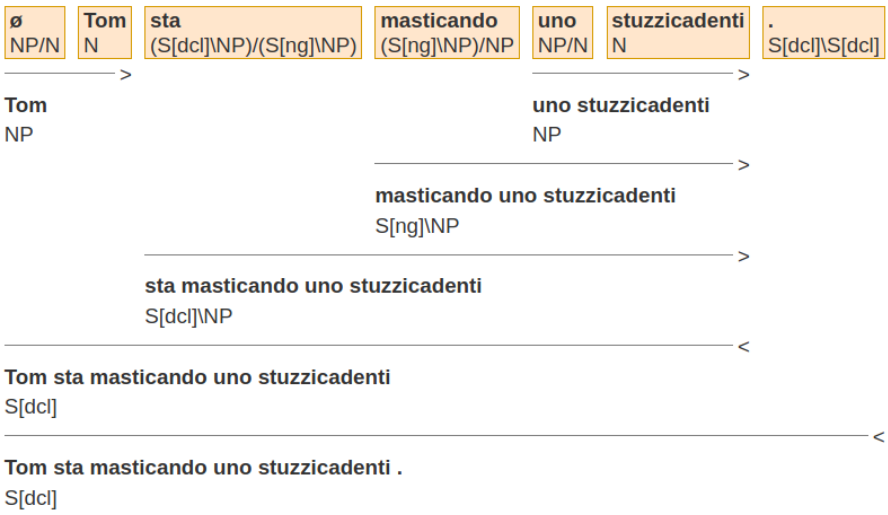


Figura 3: *Esempio di albero sintattico a costituenti che rappresenta la sintassi della frase "Tom sta masticando uno stuzzicadenti", secondo la Combinatory Categorial Grammar.*

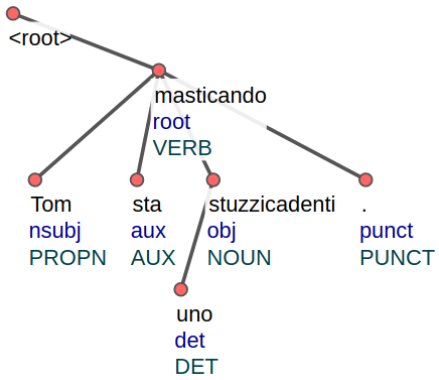


Figura 4: *Albero di dipendenze per la frase "Tom sta masticando uno stuzzicadenti", prodotto da UDpipe.*

gnificato di un'espressione in linguaggio naturale è quello delle parole e del loro significato individuale, ovvero la `semantica lessicale`.

Aprendo un dizionario, possiamo leggere parole e i loro significati, e alcune parole ne hanno più di uno. Questo fenomeno è noto come **polisemia**. Si può parlare di diversi gradi di polisemia, da parole che hanno un solo significato univoco (spesso più una parola è specifica, più tende ad avere un solo senso), fino a parole che possono indicare concetti completamente sconnessi tra loro (esempio: "calcio" come sport e come elemento chimico). La Tabella 3 mostra, a titolo esemplificativo, i sensi della parola "piano" che si trovano in `Wikizionario`[7]

In linguistica computazionale, per elaborare questo tipo di conoscenza sono state create risorse accessibili con strumenti informatici, ad esempio da un programma *software*. La più conosciuta di queste è *WordNet*, un dizionario elettronico inglese creato all'`Università di Princeton` [10]. In `WordNet`, parole e sensi formano un `grafo bipartito`[8]: ogni parola (lemma, più precisamente) è connesso ad uno o più sensi, ed ogni senso è espresso da una o più parole. Sono così rappresentate in maniera intuitiva relazioni come la **sinonimia** (diverse parole connesse allo stesso senso), o **polisemia** (diversi sensi connessi alla stessa parola). I sensi in `WordNet` sono rappresentati in effetti proprio come `synset`, insiemi di sinonimi, oltre ad avere una definizione e degli esempi. La Figura 5 mostra uno *screenshot* dell'interfaccia `Web` di `WordNet` 3.1[9] con i sensi del sostantivo inglese "rock".

La `WordNet` di `Princeton` non è l'unica rete semantica lessicale pubblicamente accessibile. Negli anni, sono state pubblicate molte `WordNet` in diverse lingue, con una struttura simile all'originale. Grazie all'iniziativa `Open Multilingual WordNet` [11], buona parte di queste risorse sono collegate tra loro e sono liberamente scaricabili o interrogabili tramite librerie *software* come il `Natural Language Toolkit` per `Python`[10].

[7] `https://it.wiktionary.org/wiki/piano`

[8] In Informatica, un grafo bipartito è un grafo in cui è possibile dividere i nodi in due gruppi con la proprietà che nessuna coppia di nodi all'interno dello stesso gruppo è connessa da un arco.

[9] `http://wordnetweb.princeton.edu/perl/webwn`

[10] `http://compling.hss.ntu.edu.sg/omw/`

Dominio	Definizione	Esempio
economia, tecnologia, ingegneria	insieme di regole prestabilite per condurre a termine un compito	
matematica, geometria, fisica	insieme di punti individuati dalla combinazione lineare di due vettori linearmente indipendenti applicati nel medesimo punto, le cui curvature fondamentali sono entrambe nulle	
architettura	livello di una struttura o di un edificio	sali al primo piano, dov'è la direzione
musica	sinonimo di pianoforte	mi diletto con il piano quando ho tempo
araldica	pianura	
geografia	estensione rettilinea di un terreno	

Tabella 3: *I significati del sostantivo "piano" secondo il Wikizionario.*

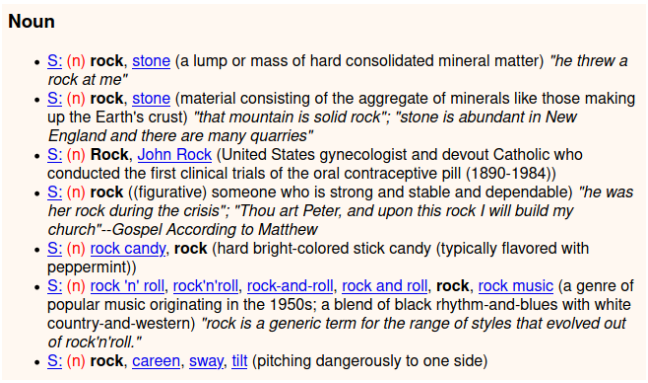


Figura 5: *Screenshot dell'Interfaccia Web di WordNet che riporta i sensi della parola inglese "rock" (sostantivo).*

Il task di attribuire il senso corretto alle parole di un'espressione in linguaggio naturale prende il nome di **disambiguazione**. È un task molto studiato in linguistica computazionale, e considerato difficile. Un algoritmo classico per la disambiguazione è l'algoritmo di `Lesk` [12], basato sulla disponibilità di un dizionario, come ad esempio `WordNet`. L'algoritmo di `Lesk` funziona nel seguente modo: data una parola e le definizioni dei suoi sensi, per ognuno di questi di conta `il numero di parole in comune` tra la definizione e il resto della frase all'interno della quale occorre la parola da disambiguare. Per esempio, considerando la parola "piano" nella frase:

> "Per trasloco dovettero spostare la centralina di tutto l'edificio ad un altro piano".

si nota come la parola "edificio" compaia sia nella frase, sia nella definizione di uno dei sensi della parola, visti nella Tabella 3, mentre non sono presenti sovrapposizioni con le altre definizioni. Si può quindi derurre che "piano" in questo contesto indica "livello di una struttura o di un edificio". L'algoritmo di Lesk è molto semplice, ma cattura l'aspetto fondamentale cruciale per la disambiguazione dei significati delle parole, ovvero che il significato dipende dal contesto. Robert Firth si riferiva proprio a questo fenomeno nello scrivere la frase citata sotto al titolo di questo articolo.

Un approccio computazionale completamente diverso al problema della semantica lessicale è quello della **semantica distribuzionale**. Questo approccio muove dalla osservazione esplicitata poco prima sul ruolo del contesto di una frase nell'attribuzione del significato alle sue parole, ed in particolare si basa fortemente sul concetto di **co-occorrenza** delle parole. Diciamo che due parole co-occorrono quando si trovano entrambe entro una distanza minima (da decidere a seconda delle applicazioni) l'una dall'altra in un testo. Una maniera per descrivere numericamente un *corpus* di testi può essere pertanto una matrice di co-occorrenze, una grande tabella in cui ogni riga e ogni colonna rappresentano una parola, e nella celle ai loro incroci viene riportato il numero di volte che si è osservata una co-occorrenza tra i due rispettivi termini. Una maniera alternativa di descrivere in termini simili un *corpus* è anche una matrice parole-documenti, in cui le righe sono sempre le parole, mentre le colonne rappresentano i documenti che compongono il *corpus*, e ad ogni incrocio si contano le occor-

Parola	D_1	D_2	D_3	D_4	D_5
Cane	1	2	1	0	0
Gatto	0	1	4	0	0
Aereo	0	1	0	1	3

Tabella 4: *Esempio di matrice di co-occorrenze parole-documenti con tre parole, cinque documenti (indicati da D_1, ..., D_5 e in ogni cella il numero di occorrenze di ogni parola in ogni documento.*

renze di ciascuna parola in tali documenti. In entrambi i casi, in corrispondenza di ogni parola, si ottiene un `vettore` numerico, una sequenza di numeri di dimensione fissa, ovvero il numero di parole nel vocabolario nel caso di matrice parole-parole o il numero di documenti nel caso di matrice parole-documenti. La Tabella 4 mostra un esempio di matrice di co-occorrenza parole-documenti. In ogni riga, si trova una parola seguita da un vettore di dimensione cinque (il numero dei documenti). Per esempio, secondo questa matrice, la parola "cane" occorre una volta nel primo documento e due volte nel secondo, mentre la parola "aereo" non occorre nel primo documenti, e così via.

Queste rappresentqazioni vettoriali delle parole sono molto utili in linguistica computazionale, poiché codificano numericamente il contesto in cui le parole occorrono, e rendono quindi verificabile computazionalmente la `ipotesi distribuzionale` [13]:

> "Parole che sono usate e occorrono in contesti simili tendono a convogliare significati simili."

In termini matematici, ci sono diversi modi di calcolare la distanza tra due vettori. Una misura molto usata per questo scopo è la **similarità coseno**:

$$K(\vec{X}, \vec{Y}) = \frac{\vec{X} \cdot \vec{Y}}{||\vec{X}||||\vec{Y}||}$$

ossia l'angolo compreso tra i due vettori, indipendentemente dal numero di dimensioni dello spazio vettoriale, normalizzato in modo da risultare un numero compreso tra -1 (vettori opposti) e 1 (vettori coincidenti). Due vettori con una alta similarità coseno sono vicini tra loro nello spazio vettoriale, in virtù del fatto di rappresentare parole che condividono un maggior numero di contesti. Secondo l'ipotesi distribuzionale, quindi, vettori con maggiore similarità coseno rappresentano parole il cui significato è vicino. Ad esempio, nel caso delle tre parole dell'esempio in Tabella 4, la similarità coseno delle rispettive coppie di parole è la seguente:

$$K(cane, gatto) \approx 0.59$$

$$K(cane, aereo) \approx 0.25$$

$$K(gatto, aereo) \approx 0.27$$

Questo strumento matematico è eccezionalmente versatile e potente, fornendo una maniera automatica e non supervisionata di calcolare rappresentazioni semantico-lessicali a partire da corpora di testo. Numerosi approcci hanno esteso questa tecnica, come ad esempio la `Latent Semantic Analysis` [14], che fa uso di `decomposizione ai valori singolari` per ridurre il numero delle dimensioni e quindi ovviare al problema della sparsità delle co-occorrenze (due parole dal significato simile potrebbero non trovarsi negli stessi contesti semplicemente perché più rare). In tempi più recenti si è iniziato a parlare di *word embedding* per riferirsi alle rappresentazioni vettoriali di parole, e sono stati proposti nuovi metodi basati sulle reti neurali per calcolarle, come nell'algoritmo `word2vec` pubblicato da `Google`.

Semantica della frase, Pragmatica e oltre

In questa sezione, abbiamo visto una carrellata di alcuni dei *task* più studiati in linguistica computazionale, dallo studio della morfologia fino alla semantica lessicale. I campi di applicazione della lingusitica compoutazionale non si esauriscono qui naturalmente, e una loro panoramica comprensiva necessiterebbe un volume di grandi dimensioni. Vorrei quindi presentarne solo alcuni, lasciando alla curiosità del lettore il compito di trovarne di interessanti.

Per cominciare, la semantica non è limitata alla sua accezione lessicale vista nella sezione precedente. Come per la sintassi, esistono diverse teorie e metodologie che mirano a fornire una rappresentazione formale (e quindi computazionale) del significato del linguaggio naturale. Tra queste, ricordiamo la `Discourse`

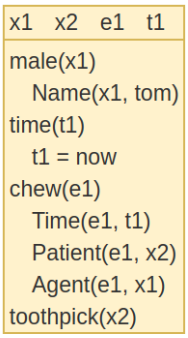


Figura 6: *Esempio di Discourse Representation Structure che rappresenta la semantica della frase "Tom sta masticando unostuzzicadenti".*

`Representation Theory` [1], che rappresenta una frase o un testo per mezzo di formule della logica del prim'ordine, o la più recente `Abstract Meaning Representation`. Il corpus multilingue `Parallel Meaning Bank`, ad esempio, è annotato con le rappresentazioni `DRT` del significato delle sue frasi (Figura 6). Un altro esempio di teoria formale della semantica è la `semantica dei frame`, dove il significato di un'espressione è caratterizzato in termini di *frame* (situazioni) e degli elementi partecipanti, ognuno col proprio ruolo.

Un gran numero di *task* sono relativi all'area chiamata `pragmatica`, ovvero lo studio della lingua dal punto di vista dell'intenzione comunicativa, e in generale di ciò che va al di là del significato in senso stretto di un'espressione in linguaggio naturale. Un *task* molto popolare sia in ricerca accademica sia in ambito industriale è l'`analisi del sentimento` (*sentiment analysis*, o *opinion mining*), che si occupa di classificare ed analizzare le opinioni soggettive e le emozioni espresse in un frammento di linguaggio naturale. Questo task è considerato cruciale anche in ambiti commerciali (es. per misurare l'interesse di clienti verso un prodotto o un servizio) e giornalistico-comunicativi (es. per conoscere l'opinione di utenti di social media), fino ad arrivare ad applicazioni ad alto impatto sociale come l'analisi automatica di cyberbullismo e hate speech.

Ci sono infine innumerevoli applicazioni che abbiamo lasciato fuori dall'esposizione per motivi di spazio, tra cui riportiamo la traduzione automatica (es. `Google Translate o DeepL`), il `question answering` e gli assistenti virtuali (es. `Alexa` o `Siri`), riassunto automatico, generazione di didascalie per immagini, e tante altre. Tutte queste applicazioni hanno come denominatore comune l'uso di una o più delle tecniche descritte in questo articolo, declinate nelle maniere più disparate per diversi obiettivi.

Apprendimento Automatico e NLP

In diversi punti in questo articolo abbiamo visto come molti task di `NLP` vengano abitualmente approcciati con metodi di apprendimento automatico, più comunemente conosciuto con il termine inglese *machine learning*. L'apprendimento automatico è una disciplina estesa e complessa di per sé, per la quale un intero articolo delle dimensioni di quello presente basterebbe solo a riassumerne le caratteristiche. Possiamo però dire che, tra le diverse famiglie di tecniche di apprendimento automatico in uso nel campo del `NLP`, la maggior parte si rifà all'apprendimento `supervisionato` (gli altri tipi principali di apprendimento automatico sono quello `non supervisionato` e il *reinforcement learning*).

Nell'apprendimento supervisionato, un modello matematico-statistico, come ad esempio una rete neurale, viene `addestrato` utilizzando una notevole quantità di `istanze`. Queste possono essere, ad esempio, frasi in linguaggio naturale o parole con associate delle etichette. Le istanze devono essere pertanto preventivamente trasformate in un formato comprensibile al calcolatore, ovvero in serie di numeri, che prendono il nome di *feature*. Nel caso del linguaggio naturale le feature possono essere molteplici: conteggi di parole in una frase, feature morfologiche, `POS`-*tag*, *word embedding*, e così via.

Il programma di apprendimento automatico ha lo scopo di creare un modello numerico che associ un certo insieme di valori delle *feature* alla corrispondente etichetta. In pratica, per addestrare un modello supervisionato si codificano le *feature* delle istanze creando un *training set* (un insieme di istanze di addestramento) e si utilizza poi un ulteriore insieme di dati (il *test set*) con le loro *feature* per misurare la qualità delle predizioni del modello. Il *training set* è tipicamen-

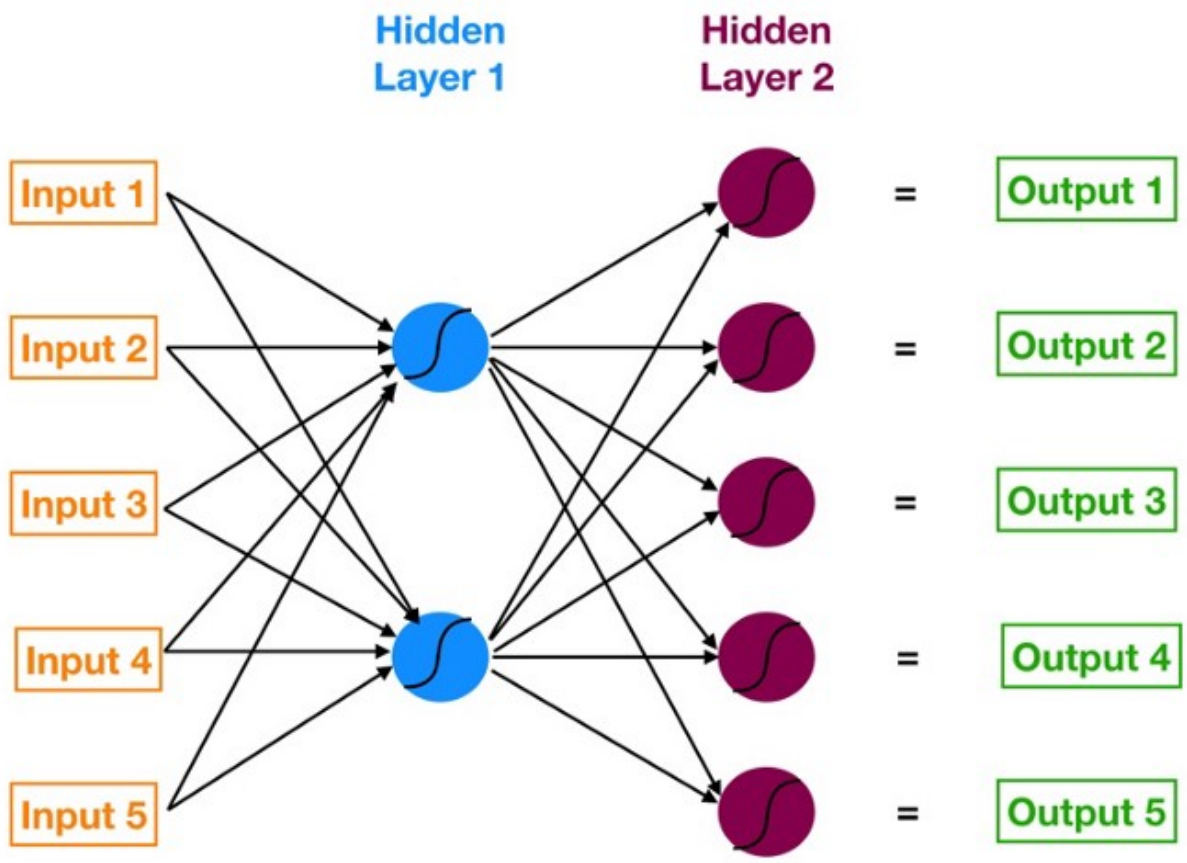


Figura 7: *Esempio di architettura di una rete neurale. Da Towards Data Science:* `https://towardsdatascience.com/understanding-neural-networks-19020b758230`

te molto più grande del *test set*, solitamente in proporzione 80%/20% oppure 90%/10%.

I modelli esistenti per apprendimento automatico sono innumerevoli, e nuovi modelli vengono presentati in continuazione da una comunità internazionale numerosa ed estremamente attiva. Tra i modelli più semplici troviamo il classificatore *naïve Bayes*, che sfrutta il noto teorema di Bayes della statistica per costruire un modello che predica la probabilità di un certo risultato a partire da una serie di osservazioni (feature). L'aggettivo *naïve* in questo caso si riferisce al fatto che questo algoritmo si basa sull'assunzione di indipendenza tra le *feature*, una assunzione relativamente `ingenua` appunto, dato che nella maggior parte dei casi le `feature` che codificano un frammento di testo hanno invece un certo grado di dipendenza reciproca.

Un altro algoritmo di apprendimento automatico supervisionato che ha conosciuto ampio successo in applicazioni `NLP` è il `Support Vector Machine`, in particolare in *task* di classificazione. Senza scendere in troppo dettaglio matematico, questo approccio vede le istanze del *training set* come dei punti in uno spazio ad alta dimensionalità, ovvero il numero di dimensioni è uguale al numero delle `feature`. Ogni istanza è quindi un punto le cui coordinate sono date dalle sue feature. L'algoritmo `SVM` poi procede iterativamente ad individuare il piano (`iperpiano` è la terminologia corretta) che separa in maniera ottimale le classi in cui sono divise le istanze.

Infine, la famiglia di metodi computazionali più usata al giorno d'oggi per task di apprendimento automatico supervisionato è quella delle `reti neurali`. Questi algoritmi, vagamente ispirati dal funzionamento del cervello umano, sono basati sulla costruzione di una rete di unità elementari di calcolo, strutturate a strati — un esempio di rete neurale è riportato in Figura 7. Le *feature* numeriche sono introdotte nel primo strato (*layer* di *input*) in maniera sequenziale, e

la computazione procede modificando ad ogni nuovo *input* i `pesi`, dei numeri che influiscono sulla computazione dei numeri contenuti nei nodi degli strati successivi. Meccanismi di apprendimento come l'algoritmo di *back propagation* fanno sì che i pesi vengano modificati in modo da ridurre l'errore rilevato tra la previsione della rete data dall'ultimo strato (*layer* di *output*) e il valore atteso che si trova nel *training set*. Le reti neurali rappresentano un ampio campo di studio, e ne esistono di moltissime forme e con diverse funzioni. Il sito `Neural Network` Zoo[11] dell'`Istituto Asimov` fornisce una interessante panoramica di alcune delle architettura più popolari.

La Valutazione dei Sistemi di NLP

Nello sviluppo di un qualunque sistema *software*, è di fondamentale importanza mettere in piedi una rigorosa e sistematica procedura di `valutazione`. Cosa vuol dire e cosa implica questa affermazione?

Valutare un sistema di `NLP` significa, in essenza, misurare la qualità della sua analisi rispetto a quello che ci si può aspettare da un essere umano medio, capace di comprendere e produrre linguaggio naturale in maniera normale. L'*output* di un sistema `NLP` va quindi confrontato con un insieme codificato di giudizi umani, un'adeguata metrica va applicata per fornire un valore numerico che indichi la somiglianza tra i due dati. Tale insieme di giudizi umano prende generalmente il nome di **gold standard** (mutuando terminologia dalla finanza).

Abbiamo già accennato ad un esempio di valutazione quantitativa nella sezione sul `POS`-*tagging*, notando come un'`accuratezza` del 99% (un valore numerico) possa considerarsi buona per una macchina ma scadente per un essere umano. Qui con il termine **accuratezza**, si intende una precisa funzione matematica: dato un insieme di possibili etichette $L = \{NN, NNP, VB, ADV, ADJ, ...\}$, una lista di n predizioni $P = \{p_1, ..., p_n\}, p_i \in L$, e un `gold standard` $G = \{g_1, ..., g_n\}, g_i \in L$,

$$\texttt{accuratezza} = \frac{\sum_{i=1}^{n} \texttt{hit}(p_i, g_i)}{n}$$

dove

$$\texttt{hit(x, y)} = \begin{cases} 1 & sex = y \\ 0 & sex \neq y \end{cases}$$

ovvero, l'`accuratezza` è il rapporto tra il numero di predizioni corrette e il numero totale di istanze da predire. Esistono molte altre metriche, per misurare diversi aspetti e applicabili a diversi *task*.

Non abbiamo ancora detto come ottenere il `gold standard`, i dati di riferimento necessari per valutare la *performance* di un sistema. A seconda del problema, ci sono diverse tecniche possibili, ma si può dire che nella maggior parte dei casi un `gold standard` si costruisce con un processo di `annotazione manuale`. Così come nel caso del `Brown corpus` a cui si è accennato in precedenza, la creazione di *corpus* annotato richiede un grande sforzo di tempo da parte di più persone. La presenza di molteplici annotazioni per ogni istanza da annotare è cruciale per garantirne la qualità, poiché persone diverse possono fornire giudizi differenti sullo stesso fenomeno, per errore, per una diversa interpretazione delle istruzioni, o per una genuina divergenza di opinioni. Una volta ottenute le annotazioni, si procede a costruire il vero e proprio `gold standard`, applicando un principio di maggioranza e in alcuni casi discutendo uno ad uno i casi in cui gli annotatori si sono trovati in disaccordo — questo processo è noto come **armonizzazione**.

Parallelamente all'armonizzazione, è importante tenere traccia della misura in cui gli annotatori si sono trovati in accordo sulle loro decisioni. A questo scopo si utilizzano misure di `inter-annotator agreement` (anche chiamato `inter-rater reliability`), una famiglia di funzioni matematiche che forniscono un valore numerico che indica il grado di accordo tra i giudizi dati su un insieme di dati. La letteratura in materia di `inter-annotator agreement` è ampia, ma nel campo del `NLP` tra le misure più usate troviamo la `Kappa di Cohen`, una misura di agreement tra due annotatori, e la `Kappa di Fleiss`, una generalizzazione che copra anche il caso in cui gli annotatori siano in numero supe-

[11] https://www.asimovinstitute.org/neural-network-zoo/

riore a due. La particolarità delle misure Kappa è che nella loro definizione tengono conto della possibilità che gli annotatori abbiano espresso lo stesso giudizio casualmente. Le misure di agreement sono importanti, e solitamente riportate nella descrizione di un *corpus* annotato, anche perché in un certo senso forniscono un limite superiore alla *performance* che ci si può aspettare da un sistema automatico. Una misura di agreement bassa è infatti un segnale che il *task* a cui sono sottoposti gli annotatori è difficile, e pertanto non è ragionevole richiedere ad un algoritmo una *performance* elevata sul medesimo *task*.

Come alternativa al processo lungo e costoso di annotazione manuale, in anni recenti sono state proposte piattaforme *on-line* tramite le quali si può sottomettere un insieme di dati ed ottenere un gran numero di annotazioni da parte di partecipanti in tutto il mondo, con una spesa relativamente contenuta. Tali piattaforme, di cui la più conosciuta è Mechanical Turk di Amazon[12] agiscono da intermediario tra il ricercatore e i partecipanti che vengono pagati tipicamente pochi centesimi per ogni istanza annotata. Questa alternativa, chiamata **crowdsourcing** allevia il problema dei lunghi tempi di annotazione da parte di un piccolo gruppo di esperti, ed è più economica dell'assunzione diretta di personale dedicato. D'altro canto, il ricercatore ha un minor controllo sulle persone che esprimono i giudizi che vanno a formare il gold standard, e pertanto si rende necessario mettere a punto procedure per garantire la qualità e la buona fede delle annotazioni.

L'annotazione manuale, sia in forma di crowdsourcing sia come annotazione da parte di esperti, è la metodologia più comune per creare gold standard per *task* di linguistica computazionale, ed è stata studiata sotto molti punti di vista. Recentemente, alcuni studiosi si stanno interrogando sulla validità della procedura usuale di annotazione e armonizzazione, ad esempio notando come il processo di armonizzazione a maggioranza tenda a rimuovere opinioni di minoranza che possono però essere rilevanti per il fenomeno studiato [15].

Per completare il quadro, quando si valuta un sistema NLP è buona norma presentare anche i risultati di un algoritmo il più possibile semplice, che fornisca una *baseline*, un termine di paragone verso il basso per misurare l'effettiva bontà del sistema proposto. Alcuni *task* hanno *baseline* molto elevate – è noto il caso della disambiguazione, per la quale scegliere il senso più comune di ogni parola porta ad una *performance* assoluta elevata. Una buona *baseline* ed una accurata misura di agreement, insieme, forniscono l'intervallo all'interno del quale può variare il risultato della valutazione di un sistema secondo una determinata metrica, facilitandone quindi l'interpretazione.

La Linguistica Computazionale in Italia

In tutto il mondo, la ricerca in Linguistica Computazionale e in Elaborazione Linguaggio Naturale si tiene principalmente all'interno di dipartimenti universitari di Informatica, Linguistica, ma anche Ingegneria o Scienze Cognitive e altri, oltre che istituti di ricerca pubblici e privati, e aziende di ogni dimensione. L'Italia non è da meno, potendo vantare un buon numero di gruppi di ricerca su tutto il territorio, in molte Università e presso centri come il Consiglio Nazionale delle Ricerche (sede dell'Istituto di Linguistica Computazionale «A. Zampolli»[13]) e la Fondazione Bruno Kessler.

Nel 2015, si è costituita l'Associazione Italiana di Linguistica Computazionale (AILC)[14], allo scopo di promuovere e supportare le attività di ricerca e divulgazione in quest'area sul nostro territorio. Tra le sue attività, AILC organizza annualmente un convegno (Italian Conference on Computational Linguistics, CLiC-it), quest'anno alla sua settima edizione[15]. Inoltre, AILC supporta l'organizzazione della campagna di valutazione delle tecnologie del linguaggio EVALITA, che dal 2007 propone ad ogni sua edizione una serie di task invitando ricercatori, studenti e aziende a sviluppare sistemi all'avanguardia per stimolare e valutare lo stato dell'arte.

[12] https://www.mturk.com/

[13] http://www.ilc.cnr.it/it/content/istituto

[14] https://www.ai-lc.it

[15] http://clic2020.ilc.cnr.it/it/home/

Le attività della AILC si svolgono in parallelo e con frequenti punti di contatto con la Associazione Italiana per l'Intelligenza Artificiale (AIxIA)[16], stabilita nel 1988, che coordina e supporta l'attività di ricerca e divulgazione sui temi della IA in Italia. Il workshop Natural Language for Artificial Intelligence (NL4AI), quest'anno alla quarta edizione, si tiene annualmente all'interno del convegno nazionale della AIxIA, e rappresenta un punto di contatto tra le due comunità[17].

Gli atti delle conferenze sopracitate, contenenti tutti gli articoli ivi presentati, sono generalmente disponibili gratuitamente online.

Conclusione

In questo articolo si è data una panoramica dei molteplici problemi e soluzioni che si incontrano quando si applicano tecniche computazionali allo studio dei fenomeni del linguaggio naturale. Abbiamo passato in rassegna numerosi *task* relativi ai diversi livelli di analisi linguistica, con cenni alle tecniche classiche e quelle più all'avanguardia, e rilevato l'importante di valutare i sistemi NLP con criteri scientifici rigorosi.

Come detto nell'introduzione, interi campi di studio sono stati esclusi dalla trattazione. Tra questi, la generazione del linguaggio naturale, che comporta tutta una serie di problematiche peculiari, a cominciare dalla sua stessa definizione. Se per l'analisi del linguaggio naturale, tema principale di questo articolo, l'*input* del problema è sempre qualche tipo di espressione linguistica e l'*output* è qualche struttura formale, come etichette di parti del discorso, o alberi sintattici, a seconda del *task*, nel caso della generazione non è affatto stabilito da dove debba partire un programma che produca linguaggio naturale. Abbiamo anche escluso l'intero campo dello studio computazionale del linguaggio parlato, che include oltre a molti degli aspetti visti in questo articolo anche sfide tecnologiche e teoriche legate al suono e alla sua forma d'onda, all'intonazione, e così via.

La comunità italiana e internazionale che studia questi problemi è ampia e vivace, e sempre aperta a nuovi studiosi interessati ai fenomeni della lingua. Se anche uno dei lettori di Ithaca si avvicinerà così all'affascinante mondo della Linguistica Computazionale, questo articolo avrà raggiunto il suo scopo.

[16] https://aixia.it/

[17] http://sag.art.uniroma2.it/NL4AI/

[1] H.Kamp, J. van Genabith, U. Reyle: *Discourse representation theory. Handbook of philosophical logic*. Springer, Dordrecht (2011).

[2] K. Evang et al.: *Elephant: Sequence Labeling for Word and Sentence Segmentation*, Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA (2013).

[3] V. Basile: *From Logic to Language: Natural Language Generation from Logical Forms*, University of Groningen, Groningen, Netherlands (2015).

[4] E. Zanchetta, M. Baroni: *Morph-it! A free corpus-based morphological resource for the Italian language*, Proceedings from the Corpus Linguistics Conference Series, Vol. 1, University of Birmingham, Birmingham, United Kingdom (2005)

[5] F. Tamburini, M. Melandri: *AnIta: a powerful morphological analyser for Italian*, Proceedings of the Eighth International Conference on Language Resources and Evaluation, European Language Resources Association, Istanbul, Turkey (2012).

[6] G. Chrupala, G. Dinu, J. van Genabith: *Learning Morphology with Morfette*, Proceedings of the sixth International Conference on Language Resources and Evaluation, European Language Resources Association, Marrakech, Morocco(2008).

[7] L. Abzianidze, J. Bjerva, K. Evang, H. Haagsma, R. van Noord, P. Ludmann, D. Nguyen, J. Bos: *The Parallel Meaning Bank: Towards a Multilingual Corpus of Translations Annotated with Compositional Meaning Representations*, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Valencia, Spain.

[8] G. Forney: *The Viterbi Algorithm: A Personal History*, University of Southern California (2005).

[9] L. Màrquez, L. Padro, H. Rodríguez: *A machine learning approach to POS tagging*, Machine Learning, 39.1, Springer (2000).

[10] G. A. Miller: *WordNet: A Lexical Database for English*, Communications of the Association for Computing Machinery, 38-11, Association for Computing Machinery, New York, NY, USA (1995).

[11] F. Bond, R. Foster: *Linking and extending an open multilingual wordnet*, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Sofia, Bulgaria (2013).

[12] M. Lesk: *Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone*, Proceedings of the 5th annual international conference on Systems documentation, Association for Computing Machinery, New York, NY, United States (1986).

[13] Z. S. Harris: *Distributional Structure*, Word, 10(2-3) (1954).

[14] T. K. Landauer, S. T. Dumais: *A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge*, Psychological review, 104(2), American Psychological Association (1997).

[15] M. Klenner, A. Göhring, M. Amsler: *Harmonization Sometimes Harms*, Proceedings of the 5th Swiss Text Analytics Conference and the 16th Conference on Natural Language Processing, CEUR Workshop Proceedings, Zurich, Switzerland (2020).

Valerio Basile: è Ricercatore di Informatica presso il gruppo Content-centered Computing del Dipartimento di Informatica dell'Università degli Studi di Torino, e dell'Hate Speech Monitoring Lab. Ha studiato Informatica a Bologna, laureandosi con il Prof. Fabio Tamburini con una tesi sulla semantica distribuzionale, e ha conseguito il dottorato di ricerca presso l'Università di Groningen, sotto la supervisione del Prof. Johan Bos con una tesi sulla generazione del linguaggio naturale. I suoi interessi di ricerca sono nel campo dell'Intelligenza Artificiale e della Elaborazione del Linguaggio Naturale. In particolare, i suoi lavori sono soprattutto negli ambiti della semantica computazionale, analisi del sentimento e del linguaggio abusivo, generazione del linguaggio naturale, analisi dei social media, e creazione di risorse linguistiche come il Groningen Meaning Bank e TWITA. È membro dell'Associazione Italiana di Linguistica Computazionale, della Associazione Italiana per l'Intelligenza Artificiale, comitati scientifici di numerose conferenze internazionali, e revisore per diverse riviste scientifiche. Durante il picco della pandemia di COVID-19 del 2020, ha pubblicato un bollettino giornaliero su linguistica computazionale e oltre, *The Copula and the Bit*.[18]

[18] http://valeriobasile.github.io/tcatb/

Machine Learning nella Fisica delle Alte Energie

Sono stato colpito dall'urgenza di fare. Conoscere non è abbastanza; dobbiamo applicare. Essere disposti non è abbastanza; dobbiamo fare.

Leonardo da Vinci

Konstantinos Bachas *Aristotle University of Thessaloniki, Thessaloniki, Greece*

Stefania Spagnolo *Dipartimento di Matematica e Fisica "E. de Giorgi", Università del Salento*
Istituto Nazionale di Fisica Nucleare, sez. di Lecce

Intelligenza Artificiale e *Machine Learning* (ML) hanno una relazione di vecchia data con la Fisica delle Alte Energie. Le tecniche di *Machine Learning* hanno giocato un ruolo importante nell'analisi dei dati della Fisica delle Alte Energie nelle due ultime decadi. Oggi, questa relazione è anche più stretta nel contesto della prossima era del Large Hadron Collider (LHC) del CERN e i suoi esperimenti. LHC rappresenta lo strumento principe della ricerca nella fisica delle alte energie e, nello stesso tempo, una sfida che motiva lo sviluppo di metodologie potenti per sfruttare al massimo l'enorme ammontare di dati prodotti nella loro complessità, alta dimensionalità e numerosità. In aggiunta, LHC è l'esempio che illustra al meglio perché la fisica delle alte energie è sempre stata un terreno fertile per innovazioni di riferimento per la gestione dati e *Machine Learning*. In questo articolo presenterò qualche esempio di applicazio-

Artificial Intelligence and Machine Learning (ML) have a long-standing relationship with High Energy Physics. Machine Learning techniques have played an important role in the analysis of high-energy physics data over the last couple of decades. Today, this relationship is even tighter in the context of the ongoing era of the Large Hadron Collider at CERN and its experiments. The latter constitute the state-of-the-art tools to perform research in high-energy physics and at the same time a challenge to find powerful ways to exploit as much as possible the huge amounts of produced data due to their complexity, high-dimensionality and size. In addition, the Large Hadron Collider is also a very illustrative example why high-energy physics has always been an excellent ground for bench-marking innovations in data science and Machine Learning. In this article, examples of application of Artificial Intelligence and

ne dell'Intelligenza Artificiale e di tecniche di *Machine Learning* nell'analisi dei dati della Fisica delle Alte Energie al Large Hadron Collider, che dimostra la potenzialtà del *Machine Learning* nell'affrontare difficili problemi di analisi dati e nel fornire soluzioni promettenti alla sfida proposta dai moderni esperimenti di fisica delle particelle elementari.

Introduzione

La Fisica delle Alte Energie (HEP) è la branca della scienza che studia le leggi fondamentali che governano la natura e le interazioni delle particelle elementari.

La nostra comprensione attuale della fisica delle alte energie è condensata nel quadro teorico denominato Modello Standard (SM). Lo SM è una teoria quantistica di campo che è stata sviluppata più di 50 anni fa e incorpora sia la meccanica quantistica che la relatività speciale.

Nonostante il suo enorme successo, lo SM non è la teoria definitiva. Alcune questioni fondamentali non trovano spiegazione nello SM e, allo stesso tempo, gli esperimenti condotti finora non hanno fornito nessuna indicazione utile alla loro soluzione. Ciò suggerisce che ci sia una teoria più fondamentale. Ci sono diversi esempi impressionanti di queste notevoli questioni. La differenza nell'intensità tra le forze elettrodeboli e la gravità, l'origine e la composizione della materia oscura, e l'apparente asimmetria tra materia e anti-materia nell'universo. Rispondere a queste domande ed esplorare in dettaglio l'origine delle masse delle particelle elementari sono le ragioni principali che hanno motivato la costruizione di LHC al CERN vicino a Ginevra, Svizzera.

Il Large Hadron Collider

LHC è il più grande e più potente acceleratore di particelle del mondo. Esso fa collidere, nel suo tunnel circolare di lunghezza pari a 27 km, posto 100 m sottoterra, due fasci di protoni che si muovono su traettorie opposte a quasi la velocità della luce e all'energia più alta mai raggiunta dall'essere umano. Fino ad ora, LHC ha funzionato eccezionalmente bene e ha prodotto una ricca messe di risultati di fisica tra i quali il più salien-

Machine Learning techniques in High Energy Physics data analysis at the Large Hadron Collider are presented, demonstrating, in yet another scientific area, the potential of Machine Learning in addressing difficult problems in data science and providing promising solutions to the challenges of modern particle physics experiments.

Introduction

High Energy Physics (HEP) is the scientific domain which studies the fundamental laws governing nature and the interactions of elementary particles.

Our current understanding of elementary particle physics is represented by the so-called Standard Model theoretical framework. The Standard Model (SM) is a quantum field theory which is being developed for more than 50 years and incorporates consistently both quantum mechanics and special relativity.

Despite its huge success, the SM is not the ultimate theory and there are still open questions that the SM cannot accommodate and have not been addressed experimentally, suggesting that there is a more fundamental theory. There are several striking examples of such outstanding questions. The difference between the strength of electro-weak and gravity forces, the origin and composition of dark matter, and the apparent matter-antimatter asymmetry in the universe. Answering these questions and fully exploring the origin of mass of the elementary particles are the main reasons for building the Large Hadron Collider (LHC) at CERN near Geneva, Switzerland.

The Large Hadron Collider

The LHC is the biggest and most powerful particle accelerator on earth, colliding, in its 100 m underground tunnel of 27 km circumference, two counter-rotating beams of protons almost at the speed of light and at the highest energy ever reached. The LHC has so far performed exceptionally well and delivered a wealth of physics results with the highlight of the Higgs boson discovery in 2012 [1, 2, 3].

te è la scoperta, nel 2012, del bosone di Higgs [1, 2, 3].

D'altra parte, LHC è anche la sorgente più abbondante al mondo di dati scientifici in termini di quantità, ma anche di complessità. I protoni dei fasci che circolano nel tunnel di LHC raggruppati in pacchetti molto densi collidono nei quattro punti di intersezione delle traettorie dei fasci (zone di interazione) ogni 25 nano secondi. Le collisioni generano il tasso di eventi, mai ottenuto prima, di 40 milioni di eventi per secondo. Ogni collisione produce un gran numero di particelle registrate dai rivelatori degli esperimenti installati attorno alle zone di interazione che leggono $O(10^8)$ sensori di diversi tipi in ogni evento. Questi grandi tassi di collisioni sono necessari a causa della natura probabilistica delle interazioni e del gran numero di diversi stati finali che possono essere generati in una singola collisione. Eventi che generano prodotti interessanti sono estremamente rari.

Di conseguenza, la velocità con cui LHC produce dati supera di ordini di grandezza le moderne capacità di processare ed immagazzinare dati. Per questo motivo soltanto una piccola frazione di questi eventi è conservata per un'ulteriore analisi. Algoritmi elaborati, ma veloci, selezionano solo alcuni eventi ritenuti interessanti sulla base di criteri fisici riducendone il tasso a circa 1 kHz. Tuttavia, anche con questo fattore di riduzione, durante il periodo di acquisizione dati detto Run 1 di LHC (2009 - 2013), il tasso di scritttura dei dati su disco era di 1 gigabyte per secondo (Gb/s), con un picco occasionale di 6 Gb/s [4]. Nel run2 di LHC (2015-2018) la velocità con cui i dati sono stati prodotti e scritti su disco è diventata di 8 Gb/s; questo valore raddoppierà nel prossimo run ed infine quintuplicherà nel run di alta luminosità di LHC (a partire dal 2027) quando si arriverà ad immagazzinare 500 petabyte di dati ogni anno.

L'analisi di questo enorme volume di dati rappresenta una sfida alla gestione dei dati e quindi per fisica delle particelle. Essa impone di considerare il (Machine Learning) (ML) una metodologia decisiva per il raggiungimento degli obiettivi scientifici di LHC in quanto essa aiuta a ridurre la dimensione dei dati concentrando l'informazione rilevante contenuta in dati di basso livello e alta dimensionalità in una rappresentazione di

Seen from another point of view, the LHC is also the world's most abundant source of scientific data in terms of size but also in terms of complexity. The colliding beams at the LHC are grouped into bunches of protons which cross every 25 nsec and the proton-on-proton collisions yield an unprecedented event rate of about 40 million events per second. Each collision produces a large number of particles which are being recorded by the LHC's detectors installed at specific locations around the accelerator ring and read $O(10^8)$ sensors of different types. These high event rates are necessary because of the probabilistic nature of the collisions and the physics channels produced in each event. Events that result in interesting products are very rare.

This event rate is orders of magnitude beyond nowadays data processing and storage capabilities. Because of this, only a small fraction of the data is saved for further analysis. For each of the produced events from a proton-on-proton head-on collision, elaborate algorithms select the interesting events based on physics criteria which reduces the rate down to roughly 100 kHz. However, even with this reduction factor, during the so-called Run 1 data-taking period of the LHC (2009 – 2013), data storing rate was 1 gigabyte-per-second (Gb/s), with the occasional peak of 6 Gb/s [4].

During Run 2 (2015-2018) the data storage speed raised to 8 Gb/s. This value will double in the next run and, finally, it will increase by a factor of five for the High Luminosity run of LHC (starting from 2027). At that point, the data volume written on tape each year will be 500 petabyte.

The analysis of this huge volume of data and the fact that the data size is expected to double at the next LHC Run scheduled for 2021-2023, pose a challenging task for data science and particle physics and indicate that Machine Learning (ML) can have a decisive effect on the LHC scientific goals as it can improve data reduction, reducing the relevant information contained in the low-level, high-dimensional data into a higher-level and lower-dimensional space.

alto livello e bassa dimensionalità.

I rivelatori degli esperimenti a LHC

L'esperimento ATLAS [5] è uno dei due grandi rivelatori *general-purpose* progettati per sfruttare l'opportunità di indagine sulle particelle elementari e le loro interazioni ad energie della decina di TeV offerta da LHC. L'apparato è installato in una caverna sotterranea a 100 m dal suolo, ha una forma cilindrica, una copertura a 4π e le sue dimensioni sono di 25 m di altezza e 46 m di lunghezza; il peso non inferiore a $7 \cdot 10^3$ tonnellate è simile a quello della torre Eiffel. ATLAS consiste di sei diversi sottosistemi di rivelazione avvolti come strati di forma cilindrica attorno al punto di collisione con la funzione di registrare il passaggio di ogni particella prodotta nelle collisioni misurandone il momento e l'energia e, in alcuni casi, identificandone la natura. Probabilmente è uno degli apparati più complessi mai costruiti.

L'esperimento CMS [6] è un rivelatore *general-purpose* a LHC con gli stessi obiettivi scientifici di ATLAS. Il rivelatore CMS è costruito attorno ad un enorme magnete solenoidale. Il rivelatore completo è lungo 21 m, largo 15 m, alto 15 m e ha un peso di circa $14 \cdot 10^3$ tonnelate.

I rivelatori ATLAS e CMS, in ogni secondo, sono spettatori di oltre un miliardo di interazioni tra particelle di altissima energia.

The detector experiments at the LHC

The ATLAS experiment [5] is one of the two large general-purpose detectors designed to exploit the physics potential of the LHC. It is installed in a cavern 100 m below ground, it has cylindrical shape and 4π coverage and dimensions of 25 m height and 46 m length while it weights not less than $7 \cdot 10^3$ tonnes, similar to the weight of the Eiffel Tower. It consists of six different detecting subsystems wrapped concentrically in layers around the collision point to record the trajectory, momentum, and energy of particles, allowing them to be individually identified and measured and is arguably one of the most complex devices ever built.

The CMS experiment [6] is a general-purpose detector at the LHC with the same scientific goals as the ATLAS experiment. The CMS detector is built around a huge solenoid magnet. The complete detector is 21 metres long, 15 metres wide and 15 metres high and has a weight of $14 \cdot 10^3$ tonnes

Over a billion particle interactions take place in each of the ATLAS and CMS detectors every second.

La necessità di *Machine Learning* nella fisica delle alte energie

Nonostante, comunemente, si abbia l'impressione che ML sia una tecnica utilizzata da pochi anni nella fisica delle particelle elementari, in realtà, essa ha giocato un ruolo importante nell'analisi dei dati per decenni. Tuttavia, solo con il programma di LHC la fisica delle particelle ha prodotto un così vasto e complesso volume di dati da rendere il ML una risorsa di utilizzo comune e diversificata. I dati registrati dagli esperimenti ATLAS e CMS sono infatti molto numerosi e ad alta dimensionalità.

Diversi approcci di ML sono stati applicati per affrontare un'ampia varietà di problemi in HEP, dalla classificazione degli eventi, alla ricostruzione delle tracce nei rivelatori, all'identificazione di alcune particelle come elettroni, fotoni o lep-

The need for Machine Learning in HEP

Although the impression is that applications of ML in HEP is something which is going on for just a few years, in fact, ML has played an important role in the physics data analysis for decades. However, not before the LHC program commenced had a particle physics experiment provided with such complex and big volumes of data. The data recorded by the ATLAS and CMS experiments are innumerable and high-dimensional.

Several ML approaches have been applied to tackle a wide variety of problems in HEP, from event classification, detector hit reconstruction, to object identification and reconstruction by using information from various detector systems, such as electron, photon, or τ lepton identification or even to identify interesting collision

toni τ mediante l'uso combinato di informazioni provenienti da vari sistemi di rivelazione. Infine il ML è utilizzato anche in algoritmi di selezione in tempo reale per l'identificazione di processi interessanti. L'uso di gran lunga più frequente del ML è avvenuto nell'ambito del problema di classificare eventi come di rumore o di segnale.

Tradizionalmente, i fisici delle alte energie usano tecniche di analisi e di riduzione dei dati che consistono in sequenze di selezioni binarie su varie proprietà dell'evento e, successivamente, costruiscono per gli eventi selezionati la distribuzione di una singola quantità che è poi processata con una procedura statistica. Questo approccio all'analisi è spesso chiamato ***cut-based analysis***. Ad esempio, per identificare l'esistenza o meno di una nuova particella prevista da un modello fisico teorico (o anche per una particella esistente in un processo noto nello SM) si cerca un sottoinsieme dei dati (ad alta dimensionalità) da collisione protone-protone per una loro analisi successiva basata sull'informazione associata a ciascuno degli eventi selezionati. L'obiettivo è ottenere sotto-campioni selezionati in modo che gli eventi dovuti a processi che producono la nuova particella - il segnale - compaiano in numero statisticamente significativo rispetto ad altri eventi - il fondo - che hanno caratteristiche simili al segnale stesso. All'ipotesi di presenza del segnale in queste regioni dei dati dovrebbe quindi corrispondere una previsione significativamente differente rispetto a quella relativa all'ipotesi nulla (assenza di segnale), permettendo, quindi, una solida verifica statistica [8].

I criteri di selezione dei dati sono guidati principalmente da considerazioni ed intuizioni fisiche che non sono facili da estendere ad alte dimensionalità e non c'è alcuna garanzia che queste selezioni siano le più efficienti. ML viene utilizzato per affrontare questi problemi fornendo una riduzione della dimensionalità e una migliore prestazione rispetto agli approcci tradizionali.

Fino ad ora, tra i numerosi algoritmi di Intelligenza artificiale, quelli che sono stati utilizzati più frequentemente per l'analisi *off-line* di dati di fisica di particelle elementari, sono le architetture semplici come *boosted decision trees* (BDT) o *shallow Neural Networks* (NN). Oggi, invece, le applicazioni di *deep neural networks* (DNN) che hanno la potenzialità di gestire problemi di alta

events and provide triggering for the detectors to record them. Most often though ML tools have been used to classify entire events as background-like or signal-like.

Traditionally, high energy physicists use data analysis and data reduction techniques utilising sequences of binary selections on the distribution of a single observed quantity, followed by a statistical treatment of the selected data. This is often called with the term **cut-based analysis**. For example, to search for the existence or not of a new particle predicted by a theoretical physics model (or even for an existing particle process known in the SM), one tries to extract subsets of the high-dimensional collision data for further analysis based of the information associated with individual events. The ultimate goal is to select these subsets in such a way that events due to processes producing the new particle -the signal- have statistical significance over other events-the background- from processes that mimic the signal ones. The signal hypothesis in these data regions would then exhibit significantly different predictions than the null hypothesis, thus allowing for an effective statistical test [8].

The decisions on the selection of the data are driven mainly by physics considerations and insights, which are not easily extended to higher dimensions nor there is any guarantee that the selections are the most efficient ones. ML comes to address these problems, by providing dimensionality reduction and improved performance with respect to the traditional approaches.

Out of the plethora of the Artificial Intelligence algorithms, simple architectures such as boosted decision trees (BDT) or shallow Neural Networks (NN), have been most frequently used in the offline HEP analysis, while applications of deep neural networks (DNN) which could adeptly handle higher-dimensional and more complex problems than previously feasible, are now more

dimensionalità più complessi rispetto a quelli affrontabili in precedenza, riscuotono sempre più interesse, anche se il loro uso non è ancora così ampiamente diffuso. Tecniche di ML sono già state applicate a varie analisi condotte sui dati degli esperimenti ATLAS e CMS qui in seguito verranno presentate due applicazioni interessanti: una riguardante l'uso di una NN ricorsiva e l'altra basata su una DNN parametrica.

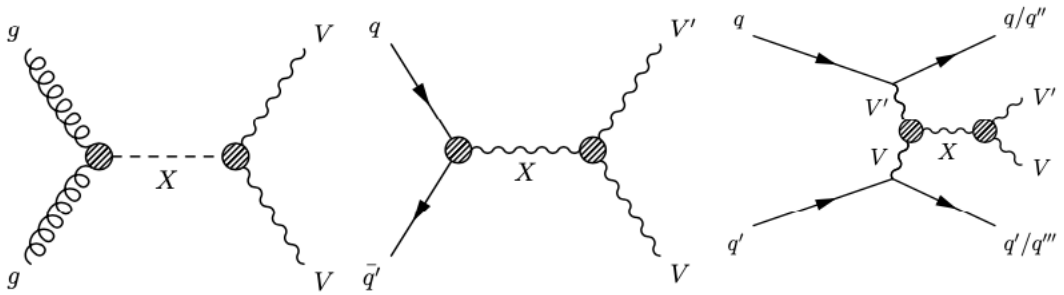
Applicazione di *recurrent Neural Networks* nella ricerca di risonanze pesanti con il rivelatore ATLAS a LHC

Molte estensioni dello SM prevedono l'esistenza di nuove particelle pesanti (risonanze) che decadono in coppie di bosoni (dibosoni) chiamati W e Z che, a loro volta, decadono in quark o leptoni (elettroni, muoni e tau, oltre agli sfuggenti neutrini). La produzione di queste coppie avviene dopo la collisione di due protoni e, più precesamente, per il tramite dell'interazione di due dei costituenti del protone, due gluoni o due quark. A seconda del modello teorico di riferimento, il dibosone previsto può essere prodotto da tre meccanismi distinti: la fusione gluone-gluone (ggF), il processo di Drell-Yan (DY) oppure la fusione vettore-bosone (VBF).

and more of gaining interest, even if they are not yet widely spread in this domain. ML techniques have been applied already to several physics analyses by both the ATLAS and CMS experiments and two interesting applications of recurrent NN and parameterized DNN are presented in the following.

Applications of recurrent Neural Networks in searches for heavy resonances with the ATLAS detector at the LHC

Many extensions to the SM predict the existence of heavy new particles (resonances) that decay into pairs of bosons (dibosons) named W and Z which in turn decay to quarks or leptons (electrons, muons, taus, besides the elusive neutrinos). The production of such pairs results from the collision of two protons and more specifically proceeds through the interaction of two of the constituents of the proton, either two gluons or two quarks. Depending on the underlying theoretical model, the predicted dibosos can be produced through three distinct mechanisms namely gluon-gluon fusion (ggF), Drell-Yan (DY), or vector-boson fusion (VBF).



(a) gluon–gluon fusion (b) Drell–Yan (c) vector-boson fusion

Figura 1: *Diagrammi di Feynman che rappresentano la produzione di nuove particelle pesanti indicate con X con i loro decadimenti in una coppia di bosoni, indicati collettivamente come V: (a) fusione gluone-gluone, (b) Drell - Yan, (c) fusione vettore-bosone. I cerchi tratteggiati rappresentano gli accoppiamenti di X con le altre particelle, sia diretti che effettivi.*
Representative Feynman diagrams for the production of new heavy particles denoted as X with their decays into a pair of bosons, denoted collectively as V: (a) gluon - gluon fusion, (b) Drell - Yan, (c) vector-boson fusion. The hashed circles represent direct or effective couplings of X to other particles.

Una rappresentazione grafica della fisica sottostante questi processi è data dai diagrammi di Feynman[1] che sono mostrati in Figura 1 per i tre meccanismi di produzione.

Come si può vedere nei diagrammi della Figura 1, dei tre processi di produzione, il ggF e il DY hanno gli stessi stati finali, mentre il VBF mostra la presenza di due quark addizionali nello stato finale. Bisogna notare che per la caratteristica della Cromo - Dinamica - Quantistica (QCD) i quark non sono mai osservati come particelle libere, ma producono un getto di particelle energetiche (adroni), che sono rivelati individualmente dall'apparato sperimentale e successivamente sono aggregate per ricostruire un'unica entità, un getto (*jet*). Nel caso di modalità di produzione VBF i due quark addizionali nello stato finale sono indicati come VBF *jet*. Quest'ultima caratteristica rappresenta una differenza chiave con le altre due modalità di produzione, ggF e DY, poiché la cinematica dei *jet* VBF differisce da quelle dei *jet* generati dal decadimento dei bosoni Z e W in quark. Tipicamente i VBF *jet* sono ben separati in **pseudorapitidy** e la loro massa invariante è grande. [2]

Queste proprietà cinematiche sono state usate in ricerche precedenti a LHC per separare processi di produzione VBF da processi di produzione in ggF/DY. In una pubblicazione recente [9] dell'esperimento ATLAS è stata, invece, usata una *recurrent neural network* (RNN) [10] per discriminare tra i meccanismi di produzione VBF e ggF/Dy. Questa classificazione del meccanismo di produzione basata su un RNN avviene all'inizio del flusso di analisi che prosegue con una successione di tagli secondo la tradizione di una *cut-based analysis* conduce alla decisione finale riguardo all'ipotesi che l'evento sia prodotto in un processo di segnale o di fondo. La referenza [9] presenta la ricerca di risonanze pesanti che decadono in WW, ZZ or WZ nella collisione protone - protone all'energia del centro di massa di 13 TeV ed è effettuata per stati finali in cui un bosone W o Z decade in leptoni mentre l'altro bosone W o Z decade in due *jet*.

[1] I diagrammi di Feynman sono un elemento essenziale del linguaggio della fisica delle particelle essendo un mezzo potente per illustrare e calcolare le transizioni tra stati nella teoria quantistica dei campi.

A pictorial representation of the fundamental physics underlying these processes is provided by the Feynman diagrams[1] which are shown in Figure 1 for the three production mechanisms.

As it can be seen in the diagrams in Figure 1, from the three production processes, the ggF and DY processes have the same final states while the VBF process exhibits two additional quarks in the final state. It should be noted that owning to the nature of Quantum Chromo-Dynamics (QCD) quarks are never observed as free particles, but they always give rise to a jet of energetic particles (hadrons), which are detected by the experimental apparatus and reconstructed as an individual object, a jet. In the case of the VBF production mode the two additional quarks in the final state are called VBF jets. The latter is a key difference with respect to the other two production mechanisms, ggF and DY, since the kinematics of the VBF-jets differ from those jets emerging from the Z or W boson decays to quarks. They are typically well separated in pseudorapidity[2] and usually has large dijet invariant mass.

These characteristics were used in previous searches at the LHC to separate VBF production from ggF/DY production. In a recent publication from the ATLAS experiment [9], a recurrent neural network (RNN) [10] is used to discriminate between the VBF and ggF/DY production mechanisms. This RNN based classification of the production mechanism is performed early in the analysis flow and once made, a "cut-based" analysis flow follows which provides the final decision of whether the event is a signal event or a background event. Reference [9] reports on a search for heavy resonances decaying into WW, ZZ or WZ using proton-proton collision data at a centre-of-mass energy of 13 TeV and is performed for final states in which one W or Z boson decays to leptons while the other W boson or Z boson decays to two jets.

[1] Feynman diagrams are an essential part of the language of particle physics as they are a powerful way to illustrate and calculate the transitions between states in quantum field theory.

[2] The psedorapidity is defined in terms of the polar angle θ as $\eta = -\ln\tan(\theta/2)$.

La motivazione per l'uso di questa architettura ML risiede nel fatto che le RNN permettono di processare sequenze di dati di lunghezza variabile. Esse possono quindi essere utilizzate per descrivere un evento in termini delle proprietà dei suoi *jet*, il cui numero varia da un evento all'altro. Alla descrizione dell'evento potrebbero inoltre contribuire utilizzando le proprietà delle altre particelle dello stato finale, elettroni, muoni o qualsiasi altra particella, che sia stata rivelata. Dato che il numero di particelle di ciascun tipo varia in ogni evento, la flessibilità dei RNN risulta necessaria per processare questo tipo di informazione.

In questo articolo della collaborazione ATLAS, il classificatore RNN usa l'informazione dei quadri-impulsi dei *jet* come variabili di input. La RNN è costruita con la libreria Keras [11] utilizzando la libreria `pyhton` Theano [12] come *back-end* per i calcoli matematici. Lo studio considera tre diversi modelli teorici che descrivono la produzione di nuove particelle pesanti che decadono in due bosoni. La sottostante distribuzione delle variabili utilizzate come *input* dal RNN e, quindi, la distribuzione di probabilità del risultato del classificatore dipendono dal modello utilizzato per descrivere la risonanza pesante, la sua massa e il modo di decadimento.

[2]La *pseudorapidity* è definita in termini dell'angolo polare θ come $\eta = -\ln\tan(\theta/2)$.

L'addestramento della RNN viene attuato utilizzando dati di fisica simulati con metodi Monte Carlo in cui una particella pesante di massa 1 TeV decade in due bosoni Z che poi decadono in due elettroni o muoni di carica opposta e due *jet* attraverso il processo che può essere riassunto in $X \to ZZ \to \ell\ell qq$, dove ℓ denota gli elettroni o muoni e q denota i quark e, in definitiva, i *jet* ricostruiti.

In Figura 2 si mostra la distribuzione di probabilità della risposta del classificatore ad eventi simulati VBF e ggF/DY corrispondenti alla produzione di una risonanza di 1 TeV nei tre diversi modelli teorici di segnale considerato in questa ricerca. Valori della risposta prossimi a 1 sono molto probabili per processi VBF, mentre quelli vicini a 0 sono tipici di eventi eventi ggF/DY.

La decisone sulla classificazione in base al meccanismo di produzione di ogni evento è quindi fissata da una soglia a 0.8 sulla risposta della

The motivation for the use of such a ML architecture lies in the fact that RNNs allow to process variable length sequences of data. It can therefore be used to describe an event in terms of the properties of its jets, whose number varies event by event. The event could as well be described using the properties of other final state particles which are reconstructed in the detector such as muons, electrons, or any other particle that appears in it. Because the number of particles of each type changes in each event, the flexibility of RNNs is needed to process such type of data.

In this paper by the ATLAS collaboration, the RNN uses information of the four-momenta of jets as input features. The RNN is built with the Keras [11] library using the Theano python library [12] as a back-end for the mathematical computations. The study considers three theoretical physics models that provide the production of a heavy new particle decaying to two bosons. The underlying distributions of the input features to the RNN and therefore the probability output of the classifier depends on the assumed model of a heavy resonance, its mass and on the decay mode.

The training of the RNN is preformed using Monte Carlo simulated physics data in which a heavy new particle of mass 1 TeV decays to two Z bosons and in turn the latter decay to two electrons or muons and two jets through the process which can be summarized as $X \to ZZ \to \ell\ell qq$, where ℓ denotes the electrons or muons and q denotes the quarks and therefore the reconstructed jets.

The output probability of the RNN classifier is shown in Figure 2 for simulated events from VBF and ggF/DY production of a 1 TeV resonance in the signal models considered in this search. The entries that are close to 1 in the score distribution are characterized as VBF-like events while the ones close to 0 are classified as ggF/DY-like events.

The actual decision on the type of production mechanism of each event is then imposed by a threshold at 0.8 on the resulting probability of

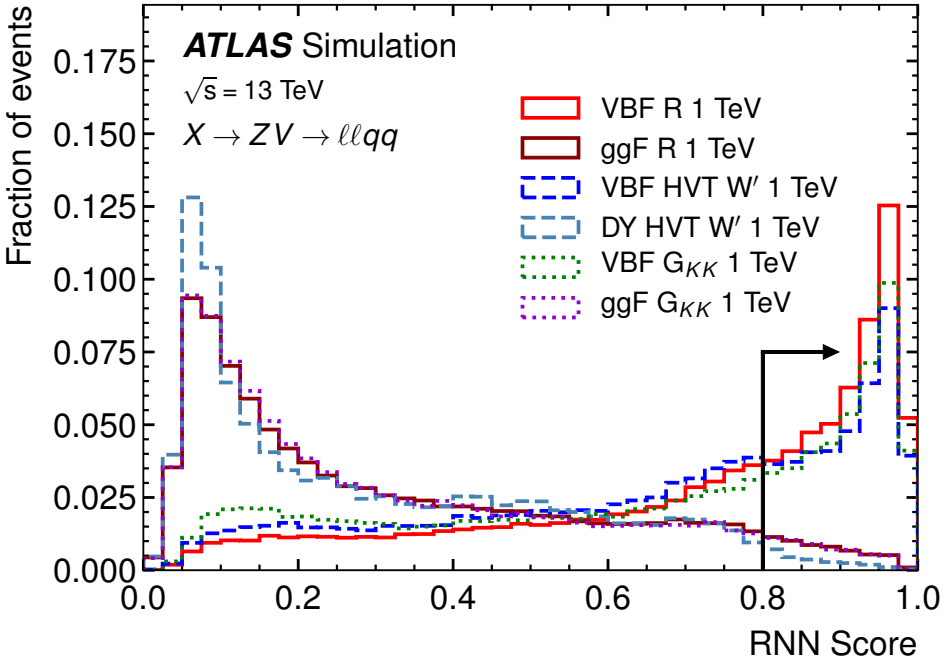


Figura 2: *Distribuzione dei risultati del classificatore RNN per la produzione di una risonanza di 1 TeV nel segnale per i modelli teorici considerati nello studio pubblicato da ATLAS [9].*
The RNN classifier score distributions for the production of a 1 TeV resonance in the signal for the theoretical physics models considered in the ATLAS publication search [9].

RNN che definisce l'evento come prodotto attraverso il meccanismo VBF o come prodotto da ggF/DY. La scelta di usare questo particolare valore di soglia è conseguente ad un'analisi dedicata in cui i valori di soglia sono stati variati cercando quello che massimizza la sensibilità ad eventi di segnale prodotti attraverso VBF.

Un'idea sulla qualità della prestazione del classificatore RNN può essere estratta dalla Figura 3 che mostra la frazione di eventi di segnale simulati che superano la soglia di 0.8 sulla risposta del RNN in funzione della massa risonante per diversi modelli di segnale. Il RNN riesce a classificare correttamente gli eventi VBF più del 40% delle volte per una risonanza dibosonica di massa superiore a 1 TeV. È opportuno osservare che l'efficienza così ottenuta migliora il risultato precedente, basato su analisi *cut-based* dal 10% al 60% (a seconda del modello di risonanza e della massa della risonanza.)

the RNN, which defines the event as an event produced through the VBF or as an event which is produced through the ggF/DY mechanisms. The motivation to use this particular threshold value is driven by a dedicated analysis which scans the threshold values looking for the one that maximizes the sensitivity to VBF signals.

An idea about the performance of the RNN classifier can be obtained from Figure 3 which shows the fractions of simulated signal events passing the RNN requirement as functions of the resonance mass for different signal models. The RNN achieves to correctly classify VBF events more than 40% of the time for a diboson resonance heavier than 1 TeV. It is remarkable that this achieved efficiency is better by 10% up to 60% (depending on the resonance model and mass) than previous cut-based selection studies.

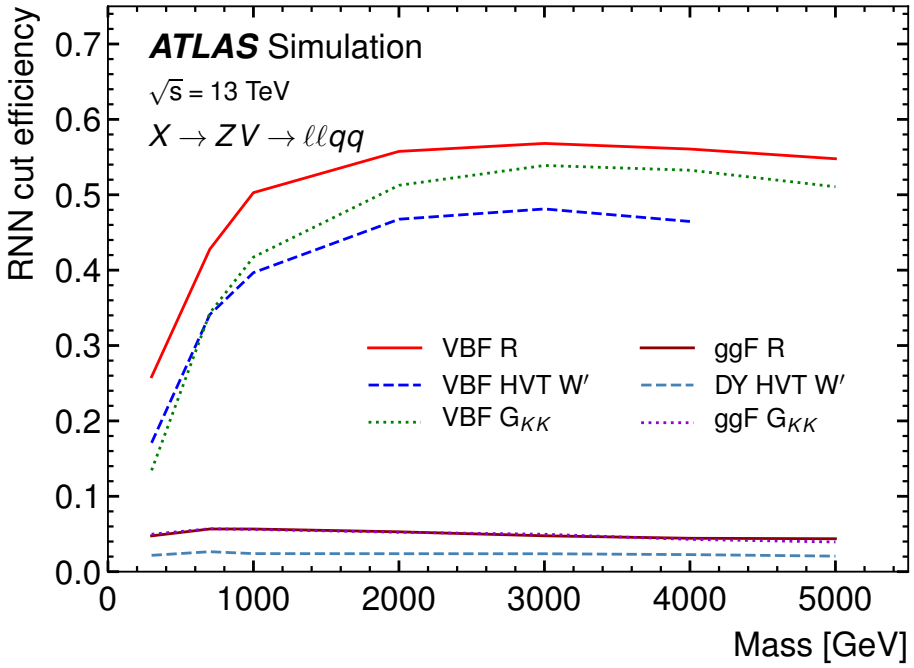


Figura 3: *La frazione degli eventi di segnale che hanno superato la soglia applicata alla risposta del classificatore RNN per definire la probabilità che l'evento sia prodotto con meccanismo VBF in funzione della massa della nuova risonanza per eventi con produzione VBF e ggF. Gli istogrammi sono estratti dalla referenza [9].*
The fractions of signal events passing the threshold on the RNN classifier probability for an event to be produced through the VBF mechanism as functions of the new resonance mass for both VBF and ggF production. Histogram from reference [9].

Applicazioni di DNN parametrizzate per la ricerca della produzione di coppie di bosoni di Higgs con il rivelatore CMS

Come menzionato nelle sezioni precedenti, tecniche di ML e di reti neurali sono state utilizzate nella fisica delle alte energie per affrontare, tra gli altri, problemi di classificazione. Nella sua più semplice e più frequente applicazione si tratta del problema di classificare eventi come compatibili con processi di fondo o di segnale. Questi ultimi prevedono la produzione di una particella ipotizzata teoricamente ma per la quale la teoria non riesce a prevedere il valore esatto della massa, ma può solo indicare un intervallo di valori possibili. Un esempio è stato presentato nella sezione prececente; un'estensione dello SM prevede una nuova particella X con una massa M_X. In accordo con certe ipotesi X decade in due Z che a loro volta decadono in $pp \to X \to ZZ \to \ell\ell qq$, dove p denota i protoni che collidono. All'interno

Applications of parameterized DNNs in searches for Higgs boson pair production with the CMS detector

As mentioned in the previous sections, Machine Learning techniques like Neural Networks have been used in HEP to address, among others, classification problems. In its simplest and more frequently met form is the signal-background classification problem where the signal is a theoretically predicted particle for which however the theory cannot predict the value of its mass but only provides a range of possible mass values. An example has already been given in the previous section; an extension to the SM theory predicts a new particle X with mass M_X. According to the hypothesis, X decays to two Z bosons which in turn decay to two leptons and two jets as $pp \to X \to ZZ \to \ell\ell qq$, where p denotes the colliding protons. Within the SM theoretical framework there are several other existing

del quadro teorico dello SM ci sono altri processi che generano gli stessi stati finali ad esempio la produzione attraverso interazioni nucleari forti di due *jet* e un bosone Z che decade in leptoni attraverso il processo $pp \rightarrow Z(\ell\ell)+$ *jet* e sono quindi considerati come fondo del processo di interesse.

In questi casi i compiti di classificazione segnale/fondo e di determinazione della massa sono distinti, ma il primo dipende dall'esito del secondo. In prima istanza per affrontare questo problema è possibile addestrare un insieme di reti neurali isolate corrispondenti ciascuna a un diverso scenario di massa. Queste reti neurali non incorporano la conoscenza di contesto generale in cui diversi valori di massa sono possibili quindi, se si cerca di usare un classificatore, che è stato addestrato a riconoscere eventi relativi ad identificare $M_X = \alpha$, per identificare la classe di eventi con $M_X = \beta$ l'accuratezza è molto bassa. Questo perché queste reti non possono interpolare bene tra diversi valori di M_X.

Per poter affrontare questo problema, nella referenza [13] è stato proposto un nuovo approccio di rete neurale parametrica in cui una singola rete neurale affronta un insieme completo di obiettivi collegati. Ciò è ottenuto estendendo la lista delle informazioni di *input* includendo in aggiunta alle quantità ricostruite quando si addestra una singola rete, anche uno o più paramtetri fisici che descrivono il modello nella sua generalità, come per esempio la massa della risonanza nel caso della ricerca di una particella instabile.

Una semplice rete neurale è addrestata usando delle caratteristiche di input x che sono le variabili che descrivono l'evento, ad esempio l'impulso e l'energia dello stato finale delle particelle. Al termine del processo di addestramento, il classificatore associa il valore 1 all'insieme delle variabili x in corrispondenza delle quali è stato allenato a produrre la risposta y in modo che f sia una buona previsione di y. Si mostra nella referenza [13] che se il problema in questione fa parte di un contesto più ampio descritto da uno o più parametri θ, è possibile costruire una rete neurale clasificatrice funzione di x e di θ. Per un dato insieme di input x, una rete neurale tradizionale genera un numero reale $f(x_0)$. Una rete parametrica, tuttavia, fornisce una mappa che è parametrizzata rispetto a θ e produce risultati

physics processes that also lead to the same final state e.g the production via strong nuclear interaction of two jets and one Z boson which decays to leptons thought the process $pp \rightarrow Z(\ell\ell)+$ jets and are therefore considered as a background to the process of interest.

In such cases, the tasks of classifying according to signal or background and of inferring the mass of the resonance are distinct, but the former depends on the actual value of the mass. To first approach this problem is tackled by training a set of isolated NNs corresponding to different mass scenarios. These networks do not incorporate knowledge of the larger context in terms of the particle's potential mass, therefore, if one tries to use a classifier, which was trained on $M_X = \alpha$, to predict the event class of an event with $M_X = \beta$ the accuracy is very poor. This is because these networks cannot interpolate smoothly between the different M_X.

To address this problem, a new NN approach has been proposed in reference [13] of a parameterized NN in which a single NN tackles the full set of related tasks. This is realized by simply extending the list of input features to include, in addition to the reconstructed quantities used when training a single network, also one or more physics parameters describing the wider scope of the problem, as for example the mass of the resonance in the resonant search case.

A plain NN is trained using as input features, x which are variables that describe the event, e.g. the momenta and energy of the final state particles. At the end of the training process, the resulting classifier provides a mapping f of x to the target value y so that f is a good predictor of the y value. If the problem in question is part of a larger context described by one or more parameters θ it has been shown in reference [13] that a NN classifier can be built that is a function of both x and θ. For a given set of inputs x, a traditional network evaluates to a real number $f(x_0)$. A parameterized network, however, provides a mapping that is parametrized with respect to θ and give different outputs values for different θ. This concept is illustrated in Figure 4.

differenti per diversi valori di θ. Questa idea è illustrata nella figura 4.

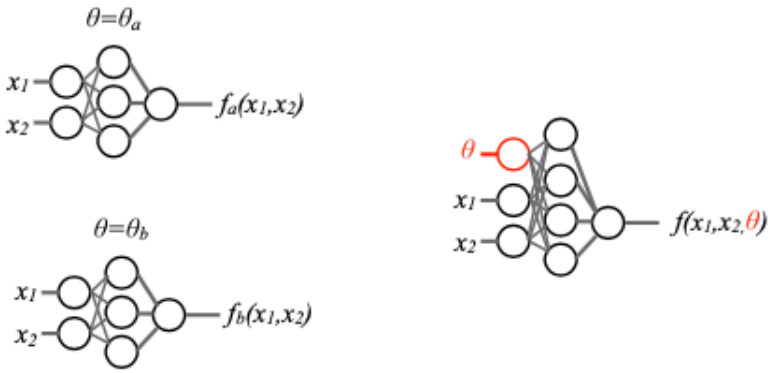


Figura 4: *Illustrazione del concetto di rete parametrica, in cui invece di addestrare network individuali con caratteristiche di input $(x_1, x_2, ..., x_n)$ per un valore fisso del parametro di θ, un solo network è addestrato a processare le informazioni $((x_1, x_2, ..., x_n, \theta)$. Da [13].*
Illustration of the concept where instead of training individual networks with input features $(x_1, x_2, ..., x_n)$ for a given fixed value of parameter θ, a single network is trained with input features $((x_1, x_2, ..., x_n, \theta)$. From [13].

Una singola rete parametrica può quindi rimpiazzare l'insieme di reti neurali individuali addestrate per specifici valori (di masse); inoltre può fornire un'interpolazione regolare della risposta in corrispondenza di valori di 0 per i quali non è stata addestrata. Questa procedura semplifica anche le difficoltà tecniche relative all'applicazione delle reti neurali perché elimina la necessità di effettuare l'addestramento (che è il passaggio più dispendioso in termini di potenza di calcolo di diverse reti) ma prevede una sola rete che gestisce tutti gli scenari di masse .

L'approccio proposto in [13] ha trovato un'applicazione diretta nell'analisi fatta dalla collaborazione CMS per la ricerca della produzione risonante e non-risonante di coppie di bosoni di Higgs (HH) in scenari oltre lo SM, coppie che decadono rispettivamente in due leptoni e due neutrini ($\ell\nu\ell\nu$, dove ℓ è un elettrone, un muone o un τ), per il tramite di un bosone W o Z, e due quark di tipo b [14]. I classificatori parametrici DNN sono stati usati per migliorare la separazione del segnale dal fondo.

In particolare, l'analisi di CMS ha usato due DNN parametriche, una per la ricerca risonante, e l'altra per quella non risonante. Nel primo caso, il parametro che gioca il ruolo di θ, come descritto sopra, è la massa della risonanza ed è fornito come input al DNN, in aggiunta alle

A single parameterized network can then replace the individual NNs trained at specific (mass) points, and also smoothly interpolate to points where it has not been trained [13]. This simplifies also the technical part of the application of the NN in such physics problems because now there is no need to train several NNs (one for each mass point) but only one network for all mass scenarios is sufficient.

The approach proposed in [13] has found direct application in the analysis performed by the CMS collaboration searching for resonant and nonresonant pair-produced Higgs bosons (HH), as extensions of the SM, decaying respectively to two leptons and two neutrinos ($\ell\nu\ell\nu$, where ℓ is either an electron, a muon, or a tau lepton), through either W or Z bosons, and two b-type quarks [14]. Parameterized DNN classifiers are used to improve the signal-to-background separation.

The CMS analysis used two parameterized DNNs, one for the resonant search and the other for the nonresonant search. In the former case, the parameter that plays the role of θ, as described above, is the mass of the resonance and is provided as input to the DNNs, in addition to

caratteristiche di *input*, che sono collegate al profilo cinematico delle particelle nello stato finale. L'insieme dei parameteri θ consiste di 13 possibili valori della massa compresi tra M_X = 260 e 900 GeV. Questi sono combinati insieme in una singola procedura di addestramento.

Nel secondo caso, i parametri θ usati sono κ_λ e κ_t: essi intervengono nella `Lagrangiana` che descrive l'interazione e modificano gli accoppiamenti del bosone di Higgs aumentando la produzione di coppie di bosoni di Higgs. Nel processo di addestramento del DNN sono state considerate 32 combinazioni di questi parametri con valori di κ_λ che variano tra -20 e 20 e κ_t tra 0.5 e 2.5.

La DNN parametrica opera tanto efficacemente quanto le DNN individuali addestrate su uno specifico valore di M_X ma necessita di un singolo processo di addestramento e produce una interpolazione continua per i casi non considerati durante la fase di addestramento. Questo è illustrato nella Figura 5. Le due curve si sovrappongono e indicano che la DNN parametrica è capace di generalizzare a casi non incontrati durante la fase di addestramento interpolando il comportamento tra punti contigui di M_X.

Le distribuzioni della variabile di uscita delle DNN per produzione risonante e non risonante di una coppia di bosoni di Higgs, dopo l'applicazione di alcuni criteri di selezione preliminari, sono riportate nella figura 6. L'accordo tra dati e predizioni dello SM è evidente negli istogrammi e indica l'assenza di nuovi fenomeni.

Queste distribuzioni di risultati da DNN parametriche sono usati come discriminanti finali in [14] e possono essere usate in ogni analisi simile per ottenere limiti sulla sezione d'urto di produzione invece di altre variabili cinematiche che sono normalmente usate come le masse invarianti ricostruite del sistema di particelle. Ciò avviene grazie al loro uso come modello per un *fit* di massima verosimiglianza finalizzato all'estrazione dell'intensità del segnale di *best fit* ossia alla determinazione della sezione d'urto. Nel caso di [14], il *fit* è fatto usando modelli costruiti dalle distribuzioni dei risultati di DNN parametriche in tre regioni della distribuzione della massa invariante dei due *jet*. Questi *fit* sono quindi usati per ottenere il limite superiore al 95 % di livello di confidenza sulla sezione d'urto di produzione per X per il processo $X \to HH \to b\bar{b}VV \to b\bar{b}\ell\nu\ell\nu$,

reconstructed input features, which are related to the kinematic profile of the final state particles. The set of θ parameters are 13 possible mass values of the resonance ranging from M_X = 260 to 900 GeV. These were combined together in a single training.

In the second case, the θ parameters used, namely κ_λ and κ_t, are parameters that enter in the Lagrangian that describes the interaction and modify the Higgs boson couplings and enhance Higgs boson pair production. 32 combinations of these parameters are considered in the DNN training with κ_λ ranging from -20 to 20 and κ_t from 0.5 to 2.5.

The parameterized DNN is able to perform as well as individual DNNs trained on specific M_X while requiring only a single training, and provides a smooth interpolation to cases not seen during the training phase. This is illustrated in Figure 5. Both curves overlap, indicating that the parameterised DNN is able to generalize to cases not seen during the training phase by interpolating the signal behaviour from nearby M_X points.

The distribution of the DNN output for the resonant and non-resonant production of the pair of Higgs bosons, after selection requirements, are shown in Figure 6. A nice agreement with the SM data is evident in the histograms and gives no hint of new phenomena.

Such parameterized DNN score distributions are then used as final discriminants in [14] and can be used in any similar analysis to obtain limits on the production cross section and replace other variables which are typically used like the reconstructed invariant mass of the system of particles. This is achieved by using them as templates to a binned maximum likelihood fit in order to extract the best fit signal cross sections. In the case of [14] the fit is performed using templates built from the parameterized DNN output distributions in three regions of the invariant mass distributions of the two jets. These fits are then used to obtain the 95% CL upper limits on the product of the production cross section for X and branching fraction for $X \to HH \to b\bar{b}VV \to b\bar{b}\ell\nu\ell\nu$, as a function of M_X.

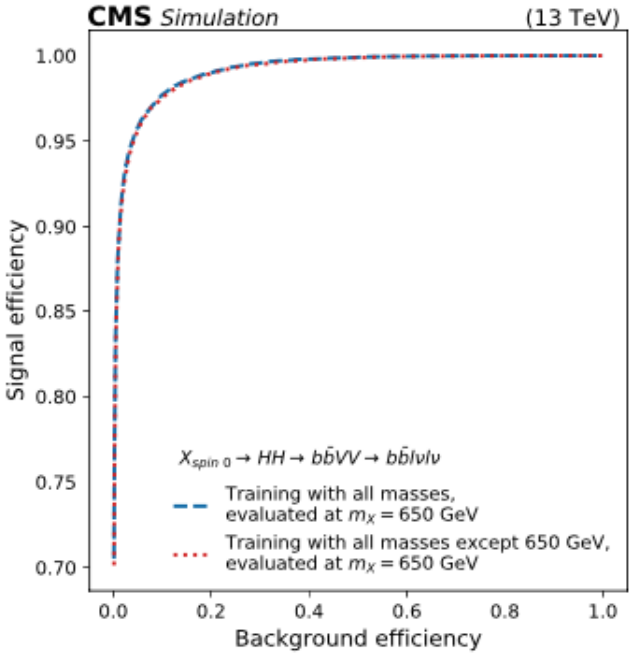


Figura 5: *Curva ROC - Receiver Operating Characteristic - dell'efficienza per il segnale in funzione della percentuale di fondo accettato come segnale per una DNN parametrica nel caso di M_X = 650 GeV. La linea tratteggiata corrisponde alla DNN addestrata su tutti i possibili segnali e calcolata per M_X = 650 GeV. La linea a punti mostra una diversa DNN addestrata su tutti i segnali tranne quello corrispondente a M_X = 650 GeV, e calcolata per M_X = 650 GeV. Istogrammi dalla referenza [14].*
ROC curve of the signal efficiency versus background efficiency for a parametrized DNN in the case where M_X = 650 GeV. The dashed line corresponds to the DNN trained on all available signal samples, and evaluated at M_X = 650GeV. The dotted line shows an alternative DNN trained using all signal samples except for M_X = 650 GeV, and evaluated at M_X = 650 GeV. Histogram from reference [14].

in funzione di M_X.

Conclusioni

Negli ultimi anni, i progressi nell'intelligenza artificiale e ML hanno permesso lo sviluppo di strumenti che hanno il potere di imprimere una nuova impronta alla strategia di analisi dati nella fisica delle alte energie.

Gli esempi di applicazione di reti neurali per affrontare ordinari problemi di classificazione a LHC presentati in questo articolo sono solo una minuscola frazione delle numerose aree dove intelligenza artificiale e ML sono applicate nella fisica delle alte energie. Nella grande maggioranza dei casi, questi approcci superano nelle prestazioni quelli tradizionali e hanno già un impatto fondamentale sulla metodologia dell'analisi dei dati che è, e sarà, effettuata nel futuro a LHC.

Conclusions

In the past few years, advances in Artificial Intelligence and Machine Learning have enabled the development of tools that have the power to shape the nature of data analysis in HEP.

The examples of the application of deep and recurrent neural networks to address common classification problems at the LHC described in this review, is only a tiny fraction of numerous areas where AI and ML are being applied in HEP. In the vast majority of the cases, these approaches outperform the traditional ones and have already a major impact on the way data analysis is being and will be conducted in the future at the LHC.

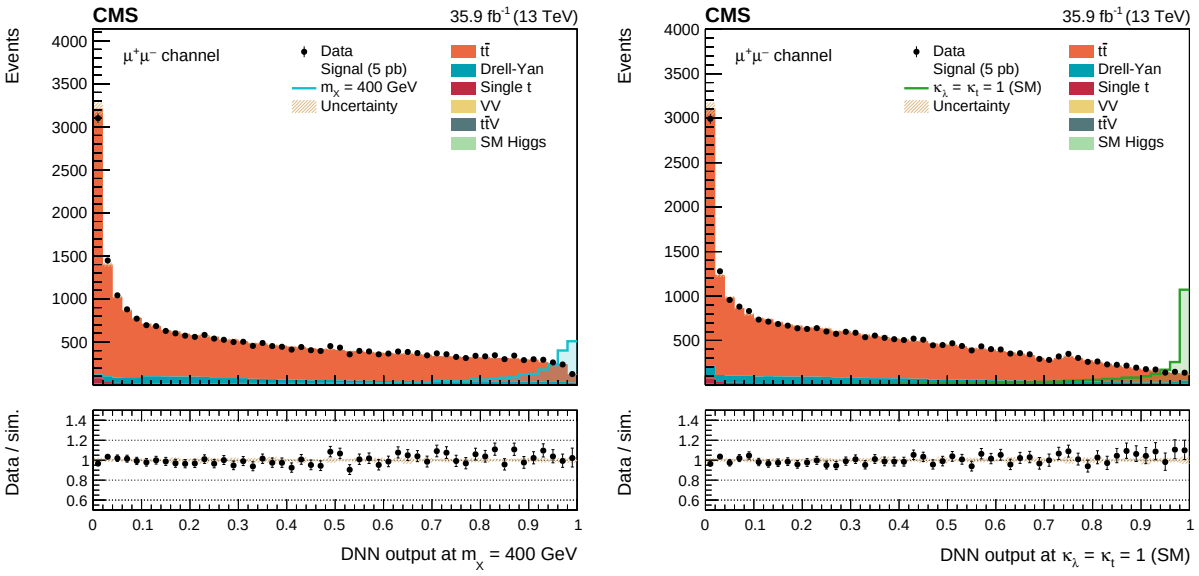


Figura 6: *(Sinistra) La distribuzione DNN ottenuta considerando dati misurati e simulati nell'analisi di [14]. I valori attorno a 0 hanno caratteristiche del fondo, quelli attorno ad 1 sono di tipo segnale. Il risultato della DNN parametrica risonante è calcolato per M_X = 400 GeV.*
(Destra) Come a sinistra per DNN parametrica non risonante calcolata con $\kappa_\lambda = 1$ e $\kappa_t = 1$.
(Left) The DNN output distribution in data and simulated events in the analysis of [14]. Output values towards 0 are background-like, while output values towards 1 are signal-like. The parameterized resonant DNN output is evaluated at M_X = 400 GeV.
(Right) As in the left panel for the parameterized nonresonant DNN with output evaluated at $\kappa_\lambda = 1$ and $\kappa_t = 1$.

[1] ATLAS collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. B 716 (2012) 1 [arXiv:1207.7214].

[2] CMS collaboration,*Observation of a new boson at a mass of* 125 *GeV with the CMS experiment at the LHC*, Phys. Lett. B 716 (2012) 30 [arXiv:1207.7235].

[3] [9] CMS collaboration, *Observation of a new boson with mass near* 125 *GeV in pp collisions at* $\sqrt{s} = 7$ *and* 8 *TeV*, JHEP 06 (2013) 081 [arXiv:1303.4571].

[4] https://wlcg-public.web.cern.ch/about

[5] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST 3 (2008) S08003.

[6] CMS collaboration, *The CMS experiment at the CERN LHC*, 2008 JINST 3 S08004.

[7] J. Neyman, E. Pearson, *On the problem of the most efficient tests of statistical hypotheses.* Phil. Trans. R. Soc. Lond. A 231, 694–706 (1933).

[8] P. Baldi, p. Sadowski, D. Whiteson, *Searching for exotic particles in high-energy physics with deep learning*. Nat Commun 5, 4308 (2014), https://doi.org/10.1038/ncomms5308

[9] The ATLAS Collaboration, *Search for heavy diboson resonances in semileptonic final states in pp collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, arXiv:2004.14636v1 [hep-ex] 30 Apr 2020, CERN-EP-2020-049, To appear EPJC.

[10] A. Sherstinsky, *Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network*, Physica D 404 (2018) 132306, arXiv: 1808.03314.

[11] F. Chollet et al., *Keras, GitHub repository (2015)*, https://github.com/fchollet/keras.

[12] R. Al-Rfou et al. (Theano Development Team), *Theano: A Python framework for fast computation of mathematical expressions*, arXiv: 1605.02688 (2016).

[13] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, D. Whiteson, *Parameterized neural networks for high-energy physics*, Eur. Phys. J. C 76 (2016) 235 [arXiv:1601.07913].

[14] The CMS collaboration, *Search for resonant and nonresonant Higgs boson pair production in the* $b\bar{b}\ell\nu\ell\nu$ *final state in proton-proton collisions at* $\sqrt{s} = 13$ *TeV*, JHEP01(2018)054.

Konstantinos Bachas: è un fisico, esperto in analisi dati, dell'Università Aristotele University di Salonicco, dove si è laureato in Fisica nel 2002. Ha ottenuto il Master of Science in fisica delle particelle elementari presso l'Università di Manchester (2003), e il dottorato di ricerca all'Università Aristotele nel 2008. È stato ricercatore presso il CERN (2009-2012), Salonicco (2013-2015), e la sezione INFN di Lecce (2016-2019). La sua attività di ricerca si è svolta principalmente nell'ambito delle attività dell'esperimento ATLAS, e ha riguardato diversi argomenti, dalla misura di processi di SM e ricerca di nuovi fenomeni nei dati di ATLAS alla del rivelatore ATLAS, con speciale attenzione allo spettrometro di muoni, e la messa a punto e verifica di sub-detector di ATLAS (rivelotori a *strip* del tracciatore interno a i tubi a *drift* dello spettrometro di muoni). Nel periodo passato a Lecce presso la sede dell'INFN è diventato esperto di tecniche moderne di ML, come DNN e RNN, che ha utilizzato per la ricerca della produzione non risonante di di-bosoni nell'analisi dei dati del run-2 di ATLAS. Il suo lavoro ha contribuito a consolidare l'uso di questi strumenti di ML come pratica comune nell'analisi dei dati di ATLAS.

Konstantinos Bachas: is a physicist and data scientist at the Aristotle University of Thessaloniki, where he graduated in Physics in 2002. He obtained a Master of Science in experimental particle physics at the University of Manchester (2003), and the Ph.D. at the Aristotle University in 2008. He was a research fellow at CERN (2009-2012), Thessaloniki (2013-2015), and INFN Lecce (2016-2019). His research activity has developed mainly within the ATLAS experiment, touching several areas, from measurements of Standard Model processes and searches for new phenomena with the ATLAS data to simulation of the ATLAS detector, with focus on the muon spectrometer, and ATLAS sub-detector commissioning and testing (strip detectors of the inner tracker and drift tubes of the muon spectrometer). During his contract with INFN, he became a crucial expert of modern Machine Learning techniques, like DNN and RNN, that he ported to the search for resonant di-boson production with the ATLAS data of run-2. His work contributed to establishing the use of these ML tools in the common practice of ATLAS data analysis.

Stefania Spagnolo: è professore associato di Fisica Nucleare e Subnucleare presso l'Università del Salento. È un componente della collaborazione ATLAS al CERN dal 2001 e ha recentemente contribuito all'esperimento PADME ai Laboratori Nazionali di Frascati (LNF) dell'INFN. Si è laureata e ha conseguito il Ph.D. in Fisica a Lecce, lavorando all'esperimento KLOE ai LNF. Dal 1998 al 2000 è stata ricercatrice del Rutherford Appletin Laboratory (UK) e componente della collaborazione OPAL a LEP (CERN). I suoi campi di interesse principali sono la fisica elettrodebole di precisione nel settore dei multibosoni e la ricerca di deviazioni dal MS che suggeriscano nuova fisica, compresi gli scenari di materia oscura non-WIMP.

Stefania Spagnolo: is professor of Nuclear and Subnuclear Physics at the University of Salento. She is working in the ATLAS experiment at CERN since 2001 and she joined in 2015 the PADME experiment at the Laboratori Nazionali di Frascati of INFN. She graduated and obtained the Ph.D. at the Università del Salento, working in the KLOE experiment at LNF. She was a research fellow at Rutherford Appleton Laboratory (UK) from 1998 to 2000 in the OPAL experiment at LEP (CERN). Her main fields of interest are tests of electroweak physics and search for deviations hinting to new phenomena connected to gauge and Higgs bosons, including scenarios for non-WIMP dark matter.

Casualità, causalità e Machine Learning nel contenimento epidemico

It is a mistake to think you can solve any major problem just with potatoes.
—— **Douglas Adams, *Life, the Universe and Everything***

Alfredo Braunstein *Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino*
Luca Dall'Asta *Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino*
Alessandro Ingrosso *The Abdus Salam International Centre for Theoretical Physics Strada Costiera 11, 34151 Trieste*

Le difficoltà principali nel contenimento di Covid-19 sono dovute ad alcune caratteristiche dell'infezione. La trasmissione del virus SARS-CoV-2 da parte di individui infetti avviene principalmente tramite l'emissione di goccioline respiratorie (*droplets*). Gli individui iniziano ad essere infettivi poco tempo dopo il contagio e sono molto spesso asintomatici[1]. Inoltre, nei casi in cui vi sia sviluppo di sintomi, questo avviene in un periodo dai 2 ai 14 giorni dopo l'infezione [2], permettendo quindi al virus di diffondersi per diversi giorni senza essere scoperto. Per poter controllare un focolaio epidemico è dunque fondamentale poter isolare i casi infetti ma privi di sintomi (asintomatici e presintomatici).

[1] La frazione di individui asintomatici è molto difficile da stimare, ma potrebbe aggirarsi attorno al 50%-75% [1]

Consideriamo il seguente esempio minimale di nascita di focolaio nell'Italia *post-lockdown*, in un periodo vicino a quello della pubblicazione di questo manoscritto:

- 19/7/2020, 19:00 - Al termine di una giornata di lavoro, Daniele e Carlo si incontrano per un breve aperitivo prima di rientrare alle rispettive abitazioni per cena.
- 20/7/2020, 16:20 - Ernesto ed il suo amico Daniele esultano abbracciandosi mentre guardano la partita della loro squadra del cuore.
- 22/7/2020, 19:45 - Amanda si prepara ad uscire per andare a trovare sua nonna Benedetta. Suo fratello Ernesto non andrà con lei: gliel'aveva preannunciato a pranzo tre giorni prima (il 19/7).
- 22/7/2020, 11:00 - Daniele si sveglia dopo una notte terribile, mal di testa fortissimo e

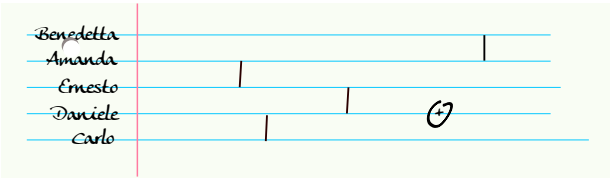


Figura 1: *Diagramma dei contatti: il tempo scorre da sinistra a destra, i segmenti verticali denotano contatti tra gli individui corrispondenti ai loro estremi. Il segno ⊕ corrisponde al momento in cui è stato prelevato il tampone, poi risultato positivo.*

febbre alta. Sarà qualcosa che ha mangiato la sera prima?

- 22/7/2020, 18:00 - Daniele decide di recarsi al pronto soccorso. Viene immediatamente sottoposto al tampone per SARS-CoV-2. Il risultato arriva dopo 40 minuti: positivo.

Il lavoro investigativo

Il giorno dopo, con in mano il risultato del test, un operatore di tracciamento (*contact tracer*) prova a ricostruire la dinamica del focolaio tramite il cosiddetto *tracciamento manuale*. Intervista Daniele, ricostruendo con lui i suoi contatti degli ultimi giorni ed determinando che ha avuto incontri che possono aver portato ad un contagio con i suoi amici Carlo ed Ernesto. Convoca dunque entrambi per effettuare dei test e (in caso di esito positivo) successiva intervista. Questo lavoro è purtroppo per sua natura lento e laborioso, e può protrarsi anche per diversi giorni. In particolare, è difficile immaginare che l'informazione possa arrivare ad Amanda in tempo per dissuaderla dal visitare Benedetta.

Supponiamo però che l'operatore tenti di portarsi avanti, svolgendo le interviste telefonicamente, senza attendere i risultati dei test. Con l'aiuto di semplice carta e penna, potrebbe poi disegnare un diagramma che rappresenta gli individui ed i loro contatti temporali, come quello riportato in Figura 1. Lo stesso operatore potrebbe, successivamente, considerare alcuni degli scenari possibili, corrispondenti a potenziali catene di infezione, a partire dai contatti avvenuti nel breve lasso di tempo costituito dai tre giorni precedenti al momento in cui è stato appreso il risultato del tampone. Per semplicità, assumerà che solo uno degli individui in questione fosse infetto il giorno 19 (siamo, dopotutto, in uno scenario *post-lockdown*, in cui i casi di infezione sono rari) e chiamerà questo individuo infetto il Paziente Zero (P0). Guardando il diagramma, l'operatore potrà subito eliminare l'ipotesi che Benedetta sia stata il paziente zero. Ecco dunque le possibilità rimaste:

Possibilità 1: Amanda è il P0: il 19 luglio contagia Ernesto, il quale a sua volta contagia Daniele durante la cena a casa sua;

Possibilità 2: Daniele è il P0;

Possibilità 3: Ernesto è il P0: il 20 luglio contagia Daniele;

Possibilità 4: Carlo è il P0: il 19 luglio contagia Daniele.

Queste possibilità potrebbero essere rappresentate come in Figura 2 (curandosi di evidenziare in ognuno dei casi tutti gli individui che potrebbero risultare infetti).

In questo piccolo esempio, tutti e cinque gli individui coinvolti possono essere stati infettati, e quattro di loro potrebbero aver ricoperto il ruolo di P0. Ogni scenario presenta possibili evoluzioni nefaste per gli attori in gioco; in particolare, Benedetta, che appartiene ad una categoria di soggetti a rischio, può essersi contagiata nell'incontro con Amanda nelle possibilità 1 e 3.

Ci troviamo in una impasse: il lavoro investigativo progressivo produce risultati puramente speculativi - per costruzione pessimistici - che possono solo fornire un'idea generica dell'estensione del focolaio. Per sua natura, inoltre, tale processo è troppo lento e non permette azioni di contenimento sufficientemente tempestive (in particolare in situazioni in cui il numero di persone coinvolte fosse significativamente maggiore rispetto all'esempio qui considerato).

Il tracciamento automatico

Un'alternativa più agile recentemente proposta è `quella del tracciamento automatico` dei contatti, basata sull'assunzione che ognuno degli individui abbia previamente installato un'applicazione (*app*) di tracciamento sul proprio *smartphone* (in Italia, `Immuni` [3]). Il *software* utilizza il segnale radio `Bluetooth Low Energy` per fornire un'ipotesi di contatto tramite una misura

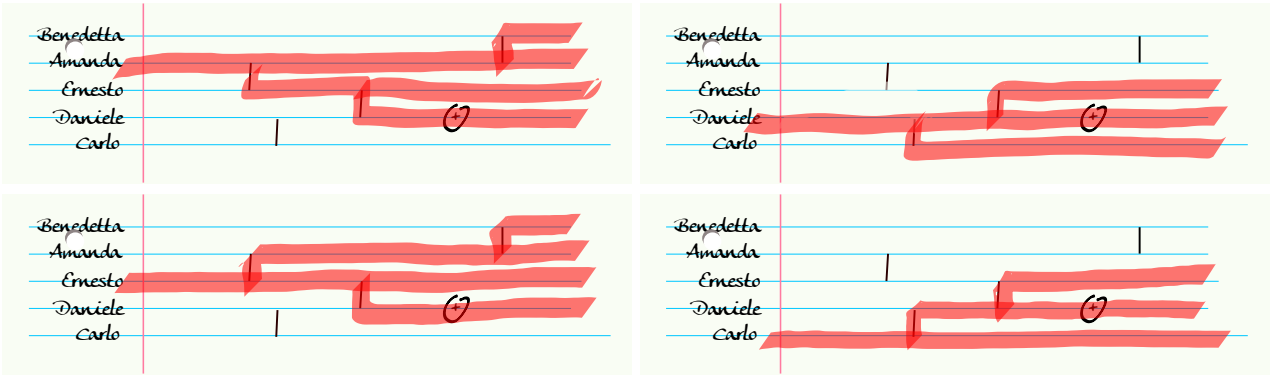


Figura 2: *Quattro delle possibili storie epidemiche (sopra: possibilità 1 e 2. sotto: possibilità 3 e 4.) che possono spiegare l'infezione di Daniele a partire dai quattro possibili pazienti zero.*

approssimata della distanza e della durata caratterizzanti la relazione di prossimità tra due dispositivi. Questo sistema permette immediatamente di evitare tutto l'iter di interviste telefoniche, avendo anche il vantaggio di rilevare contatti tra sconosciuti (ad esempio tra cliente e commerciante). Inoltre, il risultato di un test positivo viene inserito nel sistema dall'operatore, per cui Ernesto potrebbe ricevere immediatamente[2] una `notifica di esposizione`, cioè di un possibile contatto rischioso con una persona potenzialmente positiva. Parliamo di "possibile contatto" perché le caratteristiche dell'`hardware` non garantiscono grande precisione nel rilevamento della prossimità e non permettono di determinare se c'è stato veramente un contatto, e "potenzialmente positiva" semplicemente perché non è possibile sapere se Daniele fosse già positivo al momento dell'incontro. Grazie a questa notifica, Ernesto potrebbe contattare immediatamente la sua famiglia, ed Amanda potrebbe decidere, per precauzione, di non fare visita a sua nonna.

C'è un rovescio della medaglia, però: il numero di falsi allarmi generato da questo sistema può essere enorme. Se ogni individuo a cui è arrivata una segnalazione chiedesse di sottoporsi ad un test, il numero di test necessari (e la rapidità con cui questi dovrebbero poi essere processati) crescerebbe velocemente con il numero di contatti, fino a rendere il sistema stesso sostanzialmente inutilizzabile. Dovendo restringere il numero di persone da sottoporre al test, è necessario avere a disposizione un ordine o classifica di rischio d'infezione, così da sottoporre per primi al test gli individui più a rischio e solo in un secondo tempo gli altri, allorché vi sia sufficiente disponibilità di risorse per svolgere ed analizzare tutti i test.

[2]Si noti che questa descrizione non rispecchia il funzionamento di Immuni [3]. In particolare, le notifiche di esposizione in Immuni non sono immediate.

Il ruolo fondamentale della casualità

La trasmissione del virus in corrispondenza di un contatto non è un evento deterministico, così che ognuna di queste possibilità per il P0 presenterà molteplici ramificazioni: cosa è successo ad Amanda nel suo incontro con Ernesto nella Possibilità 3?

La vera natura della stima del rischio epidemico (in inglese *epidemic risk assessment*) è in realtà probabilistica. I modelli matematici considerati più accurati sono probabilistici, racchiudendo in un semplice evento stocastico (per esempio la possibile trasmissione del virus durante un contatto tra due persone) un'infinità di variabili, la maggior parte impossibili da determinare (come alcune caratteristiche del contatto e del sistema immunitario del potenziale ricevente, l'abbigliamento o eventuale protezione delle persone coinvolte, ecc.) e molte di esse ancora ignote (come sono tuttora molte delle caratteristiche della diffusione del virus SARS-CoV-2).

L'approccio più utilizzato nella letteratura scientifica per descrivere matematicamente la diffusione di Covid-19 (e più in generale di molte altre epidemie) si basa su modelli epidemici a compartimenti. In questi modelli, ogni individuo è caratterizzato ad ogni istante di tempo da uno stato interno appartenente ad un insieme finito X di possibilità (compartimenti). Nel

caso più semplice, l'individuo può essere nello stato S (sano) o I (infetto). È spesso ragionevole aggiungere almeno un terzo stato R (rimosso), per tenere conto sia della risposta immunitaria, che può rendere l'individuo immune al contagio, sia di un suo potenziale decesso; in entrambi i casi, l'individuo non è più parte del processo di diffusione del virus.

Denotiamo con $x_i^t \in X$ lo stato dell'individuo i al tempo t (assumeremo qui per semplicità t discreto, per esempio nel caso in cui si consideri una risoluzione temporale su scala giornaliera). Se denotiamo con $\partial i(t)$ l'insieme degli individui che hanno avuto contatto con i al tempo t e la configurazione dei loro stati con $x_{\partial i}^{t-1} = \{x_j^{t-1}\}_{j\in\partial i(t)}$, possiamo scrivere la distribuzione di probabilità per le traiettorie collettive del sistema $\mathbf{x} = \mathbf{x}^{0:T}$ nei tempi $0, \dots, T$ dove $\mathbf{x}^{0:t} = \{x_i^s\}_{i=1,\dots,N}^{s=0,\dots,t}$ come:

$$p(\mathbf{x}) = \prod_i p\left(x_i^0\right) \prod_{t=0}^{T-1} p\left(x_i^{t+1} | x_{\partial i}^{0:t}, x_i^{0:t}\right) , \quad (1)$$

in cui $p\left(x_i^{t+1} | x_{\partial i}^{0:t}, x_i^{0:t}\right)$ è la probabilità condizionata che l'individuo i si trovi in uno stato x_i^{t+1} al tempo $t+1$ dato lo stato suo e di tutti gli individui con cui ha avuto contatto fino al tempo t, mentre $p\left(\mathbf{x}^0\right) = \prod_i p\left(x_i^0\right)$ è la distribuzione di probabilità dello stato iniziale $\mathbf{x}^0$ del processo. Possiamo assumere di avere per quest'ultima una forma fattorizzata sugli individui, dal momento che i rari casi di infezione (pazienti zero) che danno inizio a focolai epidemici hanno solitamente origine diversa (per esempio, vengono importati da un'altra comunità) e possono pertanto essere considerati eventi indipendenti.

Prediamo come esempio il modello SIR più semplice (che chiameremo «SIR standard»), che è Markoviano in quanto lo stato di un individuo a tempo $t+1$ dipende solamente dallo stato del sistema a tempo t. Un individuo nello stato I può diventare R con probabilità μ (potremmo rappresentare concisamente questa situazione con $I \xrightarrow{\mu} R$). Inoltre, ogni individuo nello stato I può contagiare indipendentemente con probabilità λ ogni altro individuo nello stato S con cui ha un contatto (rappresentato come $IS \xrightarrow{\lambda} II$). Questo ci fornisce le seguenti espressioni

$$p(x_i^{t+1} = S | x_{\partial i}^{0:t}, x_i^{0:t}) = \delta_{x_i^t,S} \prod_{j\in\partial i(t)} (1 - \lambda\delta_{x_i^t,I})$$
$$p(x_i^{t+1} = R | x_{\partial i}^{0:t}, x_i^{0:t}) = \delta_{x_i^t,R} + \mu\delta_{x_i^t,I},$$

in cui $\delta_{a,b} = 1$ se $a = b$, e zero altrimenti. La probabilità di essere nello stato I si può ricavare analogamente o per normalizzazione. L'interpretazione di queste espressioni è semplice: nella prima equazione, per essere S al tempo $t+1$, l'individuo deve trovarsi già nello stato S al tempo t e non essere contagiato a causa dei contatti tra t e $t+1$. Nella seconda, per essere R al tempo $t+1$, o l'individuo lo era già al tempo t, oppure era I ma è stato rimosso con probabilità μ.

Il metodo Monte Carlo

Un'idea per la determinazione stocastica del rischio è quella adottata da `ViraTrace` [4], alla base dell'*app* ufficiale di tracciamento dei contatti attualmente in uso in India [5]. L'idea è semplice ed allettante. La diffusione del virus avviene nell'ombra, nascosta ai nostri occhi. Quindi, perché non replicarla in maniera visibile, usando una diffusione sintetica (che riproduca il più possibile le caratteristiche infettive del virus)? Secondo questo approccio, una volta ottenuto il risultato del test che riveli la positività di un individuo, la sua *app* inizierebbe a trasmettere l'infezione sintetica ad altre *app* di individui con cui è entrato in contatto dopo il prelievo. Data la natura stocastica del processo, ovviamente, le infezioni sintetiche possono seguire storie diverse, discostandosi tra loro e, quel che è più grave, dall'evoluzione dell'epidemia reale. La propagazione di un numero sufficientemente grande di epidemie sintetiche indipendenti permette tuttavia di esplorare gli scenari più probabili ed assegnare una probabilità di infezione ad ogni individuo (queste propagazioni possono avvenire portando avanti in parallelo $m > 1$ sistemi epidemici). A tutti gli effetti, l'approccio equivale al metodo di campionamento Monte Carlo per calcolare distribuzioni marginali $p(x_i^t)$ dalla distribuzione in (1), in cui gli individui per cui l'infezione è nota saranno considerati dei P0 al tempo del prelievo del test, e per gli individui che hanno avuto un risultato di un test negativo si assume un'immu-

nità artificiale nel periodo precedente al test (per esempio cancellando tutti i loro contatti in quel periodo). Questo tipo di campionamento non presenta particolari difficoltà e permette di produrre campioni indipendenti solo riproducendo la dinamica. Avendo a disposizione m campioni $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ distribuiti secondo (1), calcoleremo in modo approssimato il rischio infettivo dell'individuo i al tempo t come

$$p(x_i^t = I) \approx \frac{1}{m} \sum_{\mu=1}^{m} \delta_{x_i^{(\mu),t},I} \tag{2}$$

cioè la frazione di campioni in cui i risulta infetto. Un individuo che è stato sufficientemente a contatto con uno o più individui infetti noti risulterà spesso infetto nei campioni ottenuti mediante la propagazione sintetica, cosicchè anche i suoi contatti futuri saranno relativamente frequentemente infetti, e via dicendo. L'*app* di un individuo potrebbe allora quantificare questo rischio tramite una probabilità e, se la stima dovesse superare una certa soglia, potrebbe allertare il sistema sanitario nazionale oppure semplicemente suggerire al proprietario di farlo.

Questa potrebbe sembrare una soluzione non medica quasi ideale, che permetterebbe di contenere un focolaio sia di Covid-19 che di altre malattie infettive con caratteristiche simili. Purtroppo, come vedremo in seguito, ci sono diverse problematiche associate a questa soluzione. In particolare, il metodo può solo inferire infezioni da una sorgente infetta avvenute a seguito del prelievo del tampone. Dal momento che l'individuo dovrebbe essere sottoposto a quarantena immediata alla notifica di un esito positivo, l'efficacia del metodo si vanifica se l'analisi viene effettuata in tempi rapidi. In realtà, i problemi sono conseguenza del fatto che stiamo calcolando i marginali della distribuzione sbagliata!

La stima a posteriori

Nel contesto probabilistico appena descritto abbiamo sorvolato su un ingrediente fondamentale: come vengono tenuti in considerazione i risultati dei *test*? In termini Bayesiani, il risultato dei *test* ci fornisce un'evidenza $\mathcal{O}$ (ad esempio, l'individuo i è infetto al tempo t e l'individuo j è sano al tempo t'). In questo contesto, l'equazione (1) descrive una distribuzione di probabilità *a priori*, cioè precedente all'acquisizione dell'evidenza; noi siamo però interessati alla distribuzione a posteriori di $\mathbf{x} = \mathbf{x}^{0:T}$:

$$p\left(\mathbf{x}|\mathcal{O}\right) = p\left(\mathbf{x}\right) p\left(\mathcal{O}|\mathbf{x}\right) p(\mathcal{O})^{-1} \tag{3}$$

Oltre al termine in (1), ne abbiamo dunque altri due. Il termine $p(\mathcal{O})$ non è per il momento rilevante, essendo una costante che non dipende dallo stato del sistema (sarà però importante più avanti). Il termine $p\left(\mathcal{O}|\mathbf{x}\right)$ è legato alle caratteristiche dei test: per un test ideale senza difetti, è deterministico (prende valori 0 e 1) ed agisce semplicemente come un filtro, cancellando tutte quelle traiettorie collettive incompatibili con i risultati dei test; nel caso generale, attribuirà invece solo un peso minore alle traiettorie incompatibili. In ogni caso, è ragionevole assumere che i risultati dei test siano indipendenti per individui diversi, così che

$$p\left(\mathcal{O}|\mathbf{x}\right) = \prod_i p\left(\mathcal{O}_i|x_i\right) \tag{4}$$

Possiamo notare inoltre come, in presenza di evidenza di infezione (almeno un test con risultato positivo), l'ipotesi di indipendenza dello stato iniziale copra anche il caso di singolo P0: se la probabilità di essere infetto a tempo 0 tende a zero, la misura di probabilità a posteriori tenderà ad essere completamente concentrata su traiettorie con esattamente un singolo P0.

Difficoltà computazionale

Quanto è difficile stimare il rischio epidemico? In principio, basterebbe elencare tutte le possibili storie epidemiche $\mathbf{x}$ compatibili con i risultati dei test. Ad ognuna di esse è associata una probabilità secondo l'Eq. (3) (una di queste storie è quella vera, da cui la somma di tutte queste probabilità è esattamente 1). Per calcolare la probabilità a posteriori che un dato individuo sia stato infettato (o sia infetto), basterebbe mediare rispetto a questa misura:

$$p(x_i^t = I|\mathcal{O}) = \frac{\sum_{\mathbf{x}} p(\mathcal{O}|\mathbf{x}) p(\mathbf{x}) \delta_{x_i^t,I}}{\sum_{\mathbf{x}} p(\mathcal{O}|\mathbf{x}) p(\mathbf{x})} \tag{5}$$

Purtroppo, il numero di storie epidemiche $\mathbf{x}$ cresce esponenzialmente con il numero di individui,

rendendo questa soluzione proibitiva. Forse, però, non tutto è perduto: chi ci assicura che non esista una soluzione più intelligente rispetto ad elencare tutte le possibili storie epidemiche? La questione di determinare se per un dato problema esista una soluzione algoritmica praticabile è fondamentale in molti ambiti della scienza. Per la maggior parte dei problemi considerati difficili, non esiste purtroppo una dimostrazione matematica di questa difficoltà. Esiste però una vasta classe di problemi, chiamati $\mathcal{NP}$*-complete*, composta di tanti problemi notevoli (tra cui, per esempio il Problema del Commesso Viaggiatore) a cui non si è mai giunti ad una soluzione algoritmica soddisfacente (i.e. che non sia esponenzialmente lenta!), che però sono intimamente legati: se venisse scoperta (o fosse già stata scoperta) una soluzione soddisfacente per uno di essi, allora questa soluzione potrebbe essere adattata per risolvere ciascuno di questi problemi! Dimostrare che un problema appartiene a questa classe ci fornisce automaticamente una dimostrazione, se non matematica, almeno storica, della difficoltà della sua soluzione. Una soluzione efficiente di uno di questi problemi – oltre ad essere un'impresa scientifica straordinaria con enormi implicazioni – probabilmente non potrebbe rimanere nascosta a lungo: ad esempio, permetterebbe di generare un'immensa quantità di BitCoin con relativa facilità, o di superare senza difficoltà le barriere crittografiche che difendono i sistemi bancari online (il sistema RSA). Siamo dunque abbastanza convinti del fatto che una soluzione del genere non esista.

Il problema della stima del rischio epidemico non è $\mathcal{NP}$*-complete*, ma è relativamente facile dimostrare che essere capace di risolverlo efficientemente fornirebbe una soluzione efficiente anche per un problema $\mathcal{NP}$*-complete* e quindi, per tutti gli elementi della classe! Abbozzeremo qui informalmente tale dimostrazione. Il problema a cui facciamo riferimento è chiamato *Unweighted Minimum Steiner Tree* (UMST) [6]: si tratta di trovare, dato un grafo $G = (V, E)$ (dove V è l'insieme di vertici e E è l'insieme di archi) ed un sottoinsieme $Q \subset V$ di vertici (chiamati `terminali`), un sotto-albero di G che connetta tutti i terminali utilizzando un numero di archi minore di una costante predeterminata. È relativamente semplice dimostrare matematicamente (in grande generalità rispetto al modello epidemico adottato) che, per una probabilità di contagio λ sufficientemente piccola, considerando il grafo formato da tutti gli individui e dai loro contatti (che assumiamo per semplicità ripetersi nel tempo) ed un sottoinsieme Q di individui con test positivo, le storie epidemiche più probabili corrisponderanno a quelle in cui il numero di individui contagiati k è minimo e la probabilità di queste storie sarà proporzionale a λ^{k-1} (corrispondente a $k-1$ contagi). Qualunque soluzione con un numero di contagi maggiore conterrà almeno un fattore λ in più, e sarà dunque molto meno probabile. Se λ è sufficientemente piccolo, la probabilità di questo insieme di soluzioni sarà vicina ad 1! Se avessimo a disposizione un `oracolo` algoritmico efficiente per decidere se la probabilità di un individuo di essere stato infettato sia maggiore di una costante, potremmo dunque usarlo per risolvere il problema dell'UMST. Si potrebbe obiettare che non c'è ragione per cui λ debba essere così piccolo (valori ragionevoli di λ per Covid-19 potrebbero ad esempio aggirarsi tra 10^{-1} e 10^{-2} a seconda del tipo e durata del contatto) e che forse la difficoltà sia associata esclusivamente a valori estremamente piccoli e non realistici. Sostituendo però ogni arco con una catena di r archi, la probabilità che un'infezione attraversi tutta la catena risulterà un multiplo di λ^r, numero che si può rendere piccolo a piacere aumentando r. La soluzione del problema epidemico in questo grafo aumentato ci fornirebbe una soluzione del UMST nel grafo originale. In sintesi, dunque, se sapessimo risolvere efficientemente il problema della stima del rischio su un qualunque (qui, la parola "qualunque" è fondamentale) grafo di contatti, potremmo risolvere altrettanto efficientemente diversi dei problemi computazionali più difficili della storia. Questa conclusione, purtroppo, lascia ben intendere la difficoltà computazionale del problema. Ma non ci dobbiamo scoraggiare: la rete di contatti interpersonale non è una rete completamente arbitraria e non è stata disegnata con malvagità per rendere difficile la soluzione. Possiamo ancora sperare di trovare soluzioni algoritmiche, magari approssimate, che funzionino in modo accettabile nei casi reali.

L'effetto dell'evidenza

È interessante a questo punto tornare al nostro esempio per mostrare come anche esiti negativi dei test possano modificare la nostra stima a posteriori dello stato infettivo di un individuo, in modo non completamente intuitivo. Per semplicità dei calcoli, assumeremo un modello SIR standard con $\mu = 0$, i.e senza lo stato R, e con probabilità di contagio λ in ogni contatto. Calcoliamo per prima cosa la probabilità a posteriori che Benedetta sia infetta, sapendo che il test di Daniele è risultato positivo. Per far ciò, è sufficiente partire dalle storie di infezione precedenti e considerare le catene infettive che contengano Benedetta. Dopo aver diviso per la probabilità che Daniele sia infetto, otteniamo la probabilità condizionata che cerchiamo:

$$p\left(x_B^{t_1} = I | x_D^{t_2} = I\right) = \frac{2\lambda^3}{1 + \lambda^2 + 2\lambda} \qquad (6)$$

Cosa succederebbe se Carlo, dopo aver incontrato Daniele, venisse testato con risultato negativo? Seguendo la medesima logica, la probabilità a posteriori che Benedetta sia infetta è la seguente:

$$p\left(x_B^{t_1} = I | x_D^{t_2} = I, x_C^{t_3} = S\right) = \frac{2\lambda^3}{1 + \lambda^2} \qquad (7)$$

Vediamo come il risultato negativo del test eseguito su Carlo costituisca evidenza a favore del contagio subito da Benedetta! La differenza è di 2λ nel denominatore, che corrisponde alle due situazioni impossibili evidenziate nella Figura 3. Essendo λ positivo, la quantità in (7) sarà sempre maggiore di quella in (6): il fatto che Carlo non sia infetto aumenta il peso a posteriori delle storie infettive in cui Amanda o Ernesto sono i P0 e che dunque Benedetta possa essere infettata da Amanda. Si noti come questo fenomeno non possa essere assolutamente catturato dalla dinamica Monte Carlo come descritta precedentemente [4]. Quel modo di tenere conto dell'informazione acquisita permette infatti solo di influenzare gli eventi che seguono causalmente l'evidenza stessa, non consente invece di porre alcun condizionamento sulle relazioni di causalità che l'hanno preceduta temporalmente e che hanno realmente portato a tale evidenza. Così facendo, non si è in grado di migliorare la stima della probabilità di eventi correlati alla possibile

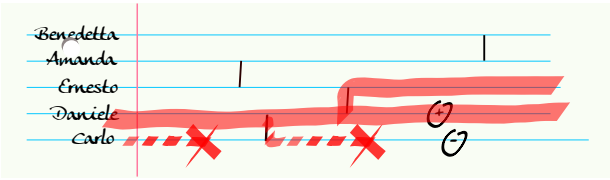


Figura 3: *L'evidenza di un test negativo di Carlo aumenta la probabilità a posteriori dell'evento che Benedetta sia infetta: confrontando* (7) *con* (6)*, due addendi λ risultano eliminati dal denominatore, corrispondenti agli eventi "Carlo è il P0 e contagia Daniele" e "Daniele è il P0 e contagia Carlo". Le infezioni di Carlo e di Benedetta sono dunque anti-correlate.*

infezione di un individuo, ma non strettamente causati da questa. Nell'esempio in questione, l'informazione addizionale del test negativo porterebbe nell'approccio Monte Carlo solo ad una diminuzione del livello di infezione degli altri individui, non potrebbe mai causarne l'aumento. Per osservare tale fenomeno, l'informazione aggiunta dall'evidenza del test negativo ha dovuto viaggiare indietro nel tempo (limitando l'insieme di possibili P0) per poi ritornare aggiornando l'infezione di Benedetta. Qualunque metodo che non permetta questo tipo di trasporto temporale dell'informazione a doppio senso non potrà mai catturare correttamente il fenomeno.

Aggiustamenti al Monte Carlo

È possibile aggiustare il metodo Monte Carlo per campionare dalla distribuzione a posteriori (3)? Una possibilità sarebbe semplicemente quella di ripesare i campioni in (2) con il termine mancante in (1) ma presente in (3):

$$p(x_i^t = I | \mathcal{O}) \approx \frac{\sum_{\mu=1}^m \delta_{x_i^{(\mu),t}, I} p(\mathcal{O}|\mathbf{x}^{(\mu)})}{\sum_{\mu=1}^m p(\mathcal{O}|\mathbf{x}^{(\mu)})} \qquad (8)$$

Questa strategia, anche se corretta, purtroppo non porta a buoni risultati perché il campionamento diventa estremamente inefficiente. Per esempio, assumendo test ideali, il termine mancante sarà diverso da zero con una frequenza esponenzialmente piccola; richiedendo un numero di campioni esponenzialmente grande per poter ottenere una stima ragionevole.

Un'altra possibilità consiste nel campionare direttamente dalla (3). In [7] viene proposto un

campionamento di Gibbs, in cui si campiona iterativamente la traiettoria di ogni singolo individuo $x_i^{0:T}$ condizionata allo stato del resto del sistema. Gli autori dello studio sostengono di ottenere buoni risultati fino a sistemi di dimensioni corrispondenti ad una piccola città (10^4 individui). Tuttavia, le proprietà di convergenza potrebbero degradarsi notevolmente al crescere della dimensione del sistema, come spesso accade nel campionamento Monte Carlo di distribuzioni complicate.

Algoritmi ispirati alla Meccanica Statistica

Considerando la dipendenza rispetto allo stato del sistema $\mathbf{x} = \mathbf{x}^{0:T}$, la distribuzione a posteriori in Eq.(3) può essere riscritta nella seguente forma:

$$p\left(\mathbf{x}|\mathcal{O}\right) = \frac{1}{Z} e^{-H(\mathbf{x})} \tag{9}$$

in cui $Z = p(\mathcal{O})$ e

$$H\left(\mathbf{x}\right) = -\log p\left(\mathbf{x}, \mathcal{O}\right) = \sum_i H_i\left(\mathbf{x}_i, \mathbf{x}_{\partial i}\right) \tag{10}$$

dove $H_i\left(\mathbf{x}_i, \mathbf{x}_{\partial i}\right) = -\log p(\mathcal{O}_i|\mathbf{x}_i) - \log p(x_i^0) - \log \prod_{t=0}^{T-1} p\left(x_i^{t+1}|x_{\partial i}^{0:t}, x_i^{0:t}\right)$. L'espressione (9) rappresenta la distribuzione di probabilità in un *ensemble* canonico (distribuzione di Boltzmann) per un modello meccanico-statistico in cui le variabili sono le traiettorie epidemiche $\{\mathbf{x}_i\}$ dei singoli individui. Ad ogni storia epidemica $\mathbf{x}$ è associata un'energia $H(\mathbf{x})$, costituita da contributi $H_i\left(\mathbf{x}_i, \mathbf{x}_{\partial i}\right)$, locali nel grafo dei contatti, cioè tali da coinvolgere la traiettoria epidemica $\mathbf{x}_i$ di un individuo i e dei suoi vicini ∂i (tutti e soli gli individui che sono venuti in contatto con esso nell'intervallo temporale preso in esame). La costante di normalizzazione

$$Z = \sum_{\mathbf{x}} e^{-H(\mathbf{x})} \tag{11}$$

rappresenta la funzione di partizione del modello. Esistono vari metodi per stimare, almeno in modo approssimato, sia la funzione di partizione che medie rispetto alla distribuzione stessa (comprese le sue distribuzioni marginali $p(\mathbf{x}_i|\mathcal{O})$, a cui siamo interessati). In particolare, mediante metodi variazionali di campo medio, si possono ottenere numericamente espressioni approssimate per queste quantità. Una possibilità è la cosiddetta approssimazione di Bethe, il cui associato schema computazionale viene chiamato algoritmo di *Belief Propagation* (BP) [8]. L'algoritmo di BP risolve una equazione di punto fisso per un vettore di dimensione elevata tramite iterazioni. Una delle caratteristiche di BP è che può essere implementato tramite dei calcoli locali, rendendo la soluzione particolarmente appetibile in un contesto distribuito: l'app di ogni individuo potrebbe in principio effettuare una parte del calcolo in cooperazione con quelle degli individui con cui è stato in contatto, evitando dunque di scambiare informazione con un server centrale. Vedremo in seguito un confronto tra diverse strategie.

Modelli ad agente

L'utilizzo pratico di procedure d'intervento basate su un approccio inferenziale richiede una validazione di tali metodi su modelli realistici di diffusione epidemica. Ad oggi, sono disponibili diversi modelli ad agenti per la simulazione su larga scala della diffusione di patogeni come il virus SARS-CoV-2 [9, 10, 11, 12, 13]. In questi modelli, ad ogni individuo è associato uno stato (p.es. Sano-Infetto-Rimosso) che evolve nel tempo, con una risoluzione temporale tipicamente giornaliera (ma alcuni modelli ammettono descrizioni a tempo continuo, p.es. [11]). L'infezione di un individuo può avvenire a causa di trasmissione diretta da parte di individui infetti o di contaminazioni ambientali, sempre però all'interno di reti di contatti realistiche, che riproducono correttamente le proprietà statistiche delle interazioni sociali, dall'ambito famigliare a quello lavorativo. Una caratteristica fondamentale di tutti i modelli realistici di diffusione del SARS-CoV-2 è il lungo periodo di incubazione e la presenza di un compartimento in cui l'agente è infettivo senza sintomi apparenti. L'analisi di dati reali ha altresì mostrato come l'infettività non sia costante nel tempo (come assunto negli usuali modelli Markoviani) ma sia una funzione del tempo con un picco a 5 giorni [9]. Ci focalizzeremo qui sul modello OpenABM, recentemente sviluppato da Ferretti et al [9], in cui 1 milione di individui interagiscono in una rete urbana a

tre differenti livelli (contatti domestici, lavorativi e casuali), la cui struttura è costruita a partire da dati demografici del Regno Unito.

Il problema del contenimento epidemico

Lo strumento fondamentale per il contenimento epidemico è quello del isolamento parziale o totale (ad es. quarantene coatte o restrizioni alla mobilità di individui o unità domestiche). Limitare il numero di persone sottoposte a quarantena è fondamentale ed in generale è auspicabile che questa sia preceduta da un risultato positivo ad un test dell'individuo interessato o di un coinquilino o familiare. Nel nostro contesto, il numero di test N_{test} che possono essere effettuati/analizzati ogni giorno è limitato (dovuto essenzialmente alla capacità di analizzarli). Quando un individuo risulta positivo viene immediatamente messo in quarantena (assumiamo per semplicità che il risultato del test sia disponibile lo stesso giorno), assieme ai componenti della sua unità domestica.

Il problema di decidere giornalmente quali individui sottoporre a test per minimizzare il numero totale (atteso) di individui infetti è un problema di *controllo stocastico*. La soluzione ottimale di problemi di controllo stocastico come questo è tipicamente estremamente complessa (ancora più difficile della valutazione del rischio, che in sostanza corrisponde al solo calcolo della funzione obiettivo!). In questo lavoro ci limiteremo dunque a proporre strategie euristiche.

Come esempi di strategie *non* probabilistiche, consideriamo due semplici metodi per la scelta degli N_{test} individui da sottoporre a test giornalmente:

- *random*: N_{test} individui scelti a caso tra quelli che non sono mai risultati positivi;
- *tracing*: gli N_{test} individui con il più alto numero di precedenti contatti con altri individui testati positivi.

La strategia `probabilistica` più diretta è quella di scegliere ogni giorno gli N_{test} individui con probabilità di infezione maggiore stimata dall'algoritmo di inferenza. Una piccola modifica, che permette di migliorarne l'efficacia, consiste nel concentrarsi sugli individui recentemente infettati (per esempio utilizzando la stima della probabilità di essere stato contagiato negli ultimi 10 giorni). Questi hanno infatti una maggiore potenzialità infettiva e si trovano vicino al «bordo» della propagazione epidemica, pertanto il loro isolamento consente di contenerne più facilmente l'avanzata. Per fortuna, la strategia probabilistica basata su Belief Propagation [8] permette anche di valutare questi aspetti più specifici del rischio infettivo.

Un confronto tra diverse strategie d'intervento in un esempio di propagazione epidemica nel modello OpenABM è riportato in Figura 4. I risultati mostrano che la stima probabilistica del rischio tramite BP (assumendo per l'inferenza un modello epidemico SIR, che è estremamente più semplice di OpenABM) consente una più accurata identificazione di casi non sintomatici (valutata dall'area sotto la ROC), risultando in un contenimento estremamente più efficace (si veda [8] per statistica e più dettagli).

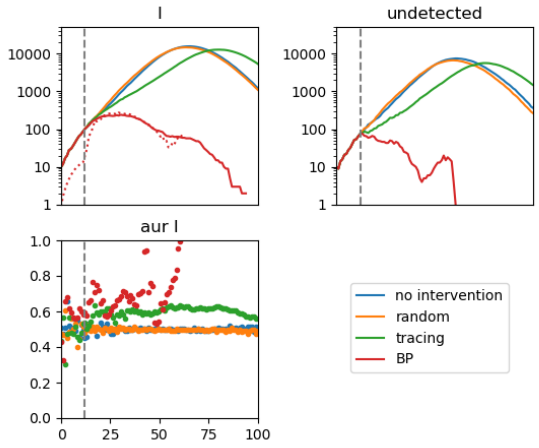


Figura 4: *Confronto tra strategie di contenimento in un esempio di una popolazione con 50K individui in cui la diffusione segue il modello OpenABM. Assumiamo che gli individui con sintomi gravi e il 50% di quelli con sintomi di media gravità vengano immediatamente isolati (o ricoverati) alla comparsa dei sintomi. Inoltre, i $N_{test} = 200$ individui selezionati dalle strategie random, tracing e BP sono sottoposti a test ogni giorno, e i casi positivi vengono isolati. L'asse orizzontale rappresenta il tempo (in giorni). Il sistema parte a tempo 0 con 10 infetti (P0), gli interventi iniziano al decimo giorno (linea tratteggiata verticale). (I): numero di individui infetti (per BP, anche la sua stima, in linea tratteggiata), (undetected): numero di infetti non testati, (aur I): area sotto la curva ROC.*

L'apprendimento automatico del modello epidemico

Nella discussione svolta sinora abbiamo trascurato un aspetto fondamentale: per essere in grado di ottenere buone predizioni, il modello utilizzato per l'inferenza deve rispecchiare il più possibile la reale dinamica di diffusione epidemica, rendendo indispensabile un processo di calibrazione dei parametri. Tale calibrazione richiede la scelta dei valori di un numero (potenzialmente molto grande) di parametri, che indicheremo congiuntamente con il vettore $\boldsymbol{\theta}$. Nel caso del modello SIR standard, ad esempio, i parametri sono solo due $\boldsymbol{\theta} = (\lambda, \mu)$: la probabilità λ che un contatto porti ad un contagio e la probabilità di guarigione istantanea μ di un individuo infetto.

Data un'osservazione parziale $\mathcal{O}$ del sistema, si vorrebbero trovare i parametri di massimo potere predittivo. Ma come misurare il potere predittivo? Definito un indice di rischio individuale – ad esempio la probabilità a posteriori di essere infetto $p(x_i = I|\mathcal{O})$ – si vorrebbe in generale valutare positivamente un metodo che assegna una stima di rischio relativamente più elevata agli individui infetti (rispetto a quelli sani). In tal senso, il potere predittivo del metodo di inferenza può essere valutato, per esempio, mediante il calcolo dell'area cumulata al di sotto della curva Receiver Operating Characteristic (ROC). Essa corrisponde infatti alla frazione di tutte le coppie di individui sano-infetto in cui il metodo di inferenza assegna un indice di rischio più alto a quello infetto. La curva ROC non può essere utilizzata direttamente per la scelta dei parametri perché utilizza dati inaccessibili (il vero stato di infezione degli individui non osservati), ma può essere utilizzata per valutare l'efficacia di altri metodi di inferenza.

Un approccio frequentemente utilizzato per l'apprendimento automatico dei parametri è quello di estendere ad essi il contesto probabilistico, assumendo che tutte le distribuzioni previamente definite siano condizionate al valore del vettore $\boldsymbol{\theta}$. Questo ci permette di definire i parametri più verosimili come quelli che massimizzano la loro probabilità a posteriori date le osservazioni $p(\boldsymbol{\theta}|\mathcal{O}) \propto p\left(\mathcal{O}|\boldsymbol{\theta}\right) p(\boldsymbol{\theta})$, oppure, in assenza di informazioni a priori per $\boldsymbol{\theta}$, semplicemente la log-likelihood

$$\mathcal{L} = \log p\left(\mathcal{O}|\boldsymbol{\theta}\right) = \log \sum_{\boldsymbol{x}} p\left(\boldsymbol{x}, \mathcal{O}|\boldsymbol{\theta}\right). \tag{12}$$

Nel contesto meccanico-statistico $-\mathcal{L}$ corrisponde alla cosiddetta energia libera del modello $-\log Z$ in (9). Come esempio dell'efficacia della massimizzazione di $\mathcal{L}$, mostriamo in figura 5 la sua approssimazione di Bethe come funzione di due parametri di un modello SIR non Markoviano, in cui il tempo di guarigione di un individuo varia con un profilo temporale caratteristico non monotono dal momento del contagio. L'inferenza è stata effettuata a partire da un insieme di osservazioni $\mathcal{O}$ su una cascata epidemica generata dal modello OpenABM (definito in [9]). Anche se l'esatto punto di massimo delle due misure non coincide, la log-likelihood (che, ricordiamo, utilizza solo i dati accessibili dell'osservazione $\mathcal{O}$) è quasi ottima in corrispondenza di un sottoinsieme di parametri che hanno massimo potere predittivo nel modello OpenABM. Il risultato è notevole, soprattutto se si tiene in considerazione che il modello probabilistico di inferenza assume una dinamica di diffusione SIR a tre stati rispetto agli 11 del modello che genera i dati [9].

La figura 5 è stata realizzata tramite il calcolo *esaustivo* di un gran numero di punti fissi di BP corrispondenti ai diversi valori combinati dei parametri. Per ognuno di essi è stata calcolata la corrispondente approssimazione di Bethe dell'energia libera. All'aumentare del numero di parametri – in un'ottica di calcolo distribuito – questa soluzione risulta ovviamente troppo lenta per essere messa in pratica. Diventa allora essenziale trovare un metodo *non esaustivo* per la ricerca dei parametri ottimali, massimizzando direttamente $\mathcal{L}$. Purtroppo, il problema della massimizzazione di $\mathcal{L}$ è per sua natura particolarmente ostico, per via della sommatoria su un numero esponenziale di stati epidemici in (12). Un approccio spesso utilizzato in questa situazione è quello di *Expectation Maximization* (EM). In EM, si sfrutta l'espressione variazionale dell'energia libera del sistema canonico

$$-\log Z = \langle H \rangle_p - S(p) \tag{13}$$

$$= \min_q \langle H \rangle_q - S(q) \tag{14}$$

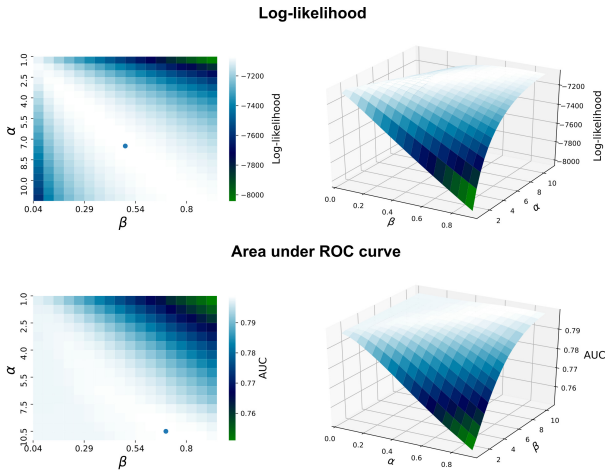


Figura 5: *Log-likelihood e area cumulata dalla curva ROC in funzione di due parametri (α, β) caratteristici della funzione temporale di guarigione (distributione gamma con forma α e scala inversa β) su un'epidemia di 10 000 individui generata da OpenABM. Si veda [14] per ulteriori dettagli.*

dove $p = p(\mathbf{x}|\mathcal{O}, \boldsymbol{\theta})$, $q(\mathbf{x})$ è una distribuzione arbitraria, $\langle\cdot\rangle_q$ denota una media rispetto alla probabilità q, ed S è l'entropia di Shannon $S(q) = -\sum_{\mathbf{x}} q(\mathbf{x}) \log q(\mathbf{x})$. Nel nostro caso specifico, massimizzando su $\boldsymbol{\theta}$ otteniamo

$$\max_{\boldsymbol{\theta}} \mathcal{L} = \max_{\boldsymbol{\theta}, q} \langle \log p(\mathbf{x}, \mathcal{O}|\boldsymbol{\theta}) \rangle_q + S(q) \qquad (15)$$

Partendo da una stima iniziale (anche rozza o banale) del vettore dei parametri $\boldsymbol{\theta}_0$, l'idea di EM è quella di risolvere la doppia ottimizzazione con una massimizzazione alternata e reiterata su q (a $\boldsymbol{\theta}$ fissato) e su $\boldsymbol{\theta}$ (a q fissato). Notando che nella k-esima iterazione l'ottimizzazione su q a $\boldsymbol{\theta}_k$ fissato comporta che q debba banalmente essere uguale a $p_k(\mathbf{x}) = p(\mathbf{x}|\mathcal{O}, \boldsymbol{\theta}_k)$ grazie a (13)-(14) e sfruttando il fatto che nell'espressione a q fissato S non dipende da $\boldsymbol{\theta}$, l'iterazione si semplifica in

$$\boldsymbol{\theta}_{k+1} = \arg\max_{\boldsymbol{\theta}} \langle \log p(\boldsymbol{x}, \mathcal{O}|\boldsymbol{\theta}) \rangle_{p_k} . \qquad (16)$$

In un punto fisso di (16), la log-likelihood è stazionaria rispetto al vettore di parametri $\boldsymbol{\theta}$. Il grosso vantaggio di EM è l'apparentemente miracoloso scambio effettivo tra logaritmo e somma (diventata media) da (12) a (16). Essendo nel nostro caso $\log p(\boldsymbol{x}, \mathcal{O}|\boldsymbol{\theta})$ una somma di termini locali (si veda Eq. (10)), la media in (16) risulterà facile da stimare con BP. Purtroppo la dipendenza dai parametri $\boldsymbol{\theta}$ è nel nostro caso ancora troppo complicata per poterla ottimizzare in modo esplicito. Un modo efficace per implementare una strategia di ottimizzazione è quello di percorrere successivamente piccoli passi nella direzione di massima crescita, usando l'informazione direzionale del gradiente $\nabla_{\boldsymbol{\theta}} \langle \log p(\boldsymbol{x}, \mathcal{O}|\boldsymbol{\theta}) \rangle_{p_k}$ – questo non è altro che il metodo della Discesa del Gradiente (o più propriamente nel nostro caso, ascesa), classicamente utilizzato in ambito di ottimizzazione convessa, ora divenuto la tecnica algoritmica standard nell'Apprendimento Automatico dei *Deep Networks*. Previo scambio di ∇ e $\langle\cdot\rangle$, questo gradiente diventa una somma di termini locali, che può essere calcolata facilmente tramite scambio di messaggi tra nodi vicini. Nella pratica, si è visto che è sufficiente alternare un singolo step nelle equazioni di BP ad un piccolo aggiornamento dei parametri, utilizzando la stima corrente del gradiente [15]. Questa procedura viene ripetuta in maniera iterativa sino al raggiungimento di un punto fisso, in corrispondenza del quale l'approssimazione di Bethe della log-likelihood è stazionaria rispetto ai parametri. Come quelli precedenti, anche questo calcolo può essere effettuato in modo distribuito tramite scambi locali tra le app di individui che sono stati in contatto.

Conclusioni

Le tecniche di inferenza probabilistica e di Machine Learning possono avere un ruolo primario nel contrasto alla diffusione dei patogeni. Il caso del virus SARS-Cov-2 è emblematico: buona parte della dinamica diffusiva è nascosta all'osservatore a causa dell'alta percentuale di individui asintomatici infettivi. Il ricorso al paradigma probabilistico pone, come si è visto, formidabili problemi computazionali, che possono essere affrontati ricorrendo a metodi di approssimazione sviluppati nel contesto della Meccanica Statistica. L'inferenza su modelli stocastici a compartimenti consente di individuare soggetti ad alto rischio di infezione e di guidare procedure di test e quarantene. Inoltre, l'evidenza dei test, per sua stessa natura parziale e rumorosa, permette di acquisire importanti informazioni sulla meccanica stessa della diffusione del virus in popolazione, attraverso una costante ri-calibrazione automatica del modello inferenziale. Per concludere, fac-

ciamo notare che la procedura algoritmica qui descritta – essendo basata su scambio di messaggi tra individui e non prevedendo il ricorso ad un'unità centralizzata di elaborazione ed immagazzinamento dati – risulta un ottimo candidato per lo sviluppo di un sistema scalabile ed attento alla privacy degli individui. Avvertiamo tuttavia il lettore che gli argomenti discussi riguardano solo aspetti epidemiologici e che uno sviluppo concreto di questi sistemi presenta sicuramente altre sfide di carattere tecnologico.

Ringraziamenti

Ringraziamo vivamente i nostri amici e colleghi del Politecnico di Torino I. Biazzo, G. Catania, F. Mazza e A.P. Muntoni che hanno svolto con noi la ricerca sugli argomenti qui presentati (in particolare A.P.M. per la Figura 5). Inoltre, A.B. ed L.D. ringraziano F. Ricci, E. Marinari e L. Ferretti per interessanti discussioni.

[1] M. Day: *Covid-19: identifying and isolating asymptomatic people helped eliminate virus in italian village.*, BMJ, 368 (2020) m1165.

[2] Centers for disease control and prevention. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html. Aggiornato a Settembre 2020.

[3] Immuni contact tracing app. https://www.immuni.italia.it/, 2020.

[4] I. Bestvina. Viratrace. https://github.com/ViraTrace/InfectionModel, 4 2020.

[5] L. Simmonds. Could croatian startup rescue tourism from coronavirus consequences? https://www.total-croatia-news.com/made-in-croatia/43538-coronavirus, (2020)

[6] M. R. Garey and D. S. Johnson.: *Computers and Intractability: A Guide to the Theory of NP-Completeness (Series of Books in the Mathematical Sciences)* W. H. Freeman, New York (1979).

[7] R. Herbrich, R. Rastogi, R. Vollgraf: *Crisp: A probabilistic model for individual-level Covid-19 infection risk estimation based on contact data*, arXiv:2006.04942 (2020).

[8] A. Baker at al. : *Epidemic mitigation by statistical inference from contact tracing data, 2020*, arXiv:2009.09422.

[9] L. Ferretti et al.: *Quantifying SARS-Cov-2 transmission suggests epidemic control with digital contact tracing*, Science, 368 (2020) eabb6936.

[10] M. Chinazzi et al.: *The effect of travel restrictions on the spread of the 2019 novel coronavirus (Covid-19) outbreak*, Science, 368 (2020) 395.

[11] L. Lorch at al. : *A spatiotemporal epidemic model to quantify the effects of contact tracing, testing, and containment*, arXiv:2004.07641 (2020).

[12] C. C. Kerr at al. : *Covasim: an agent-based model of Covid-19 dynamics and interventions.* medRxiv (2020). https://doi.org/10.1101/2020.05.10.20097469

[13] J. A. Moreno López et al. : *Anatomy of digital contact tracing: role of age, transmission setting, adoption and case detection*, medRxiv (2020) https://doi.org/10.1101/2020.07.22.20158352

[14] Sibyl website. https://github.com/sibyl-team (2020).

[15] A. Braunstein, A. Ingrosso: *Inference of causality in epidemics on temporal contact networks*, Sci. Rep, 6 (2016) 27538.

Alfredo Braunstein: è Professore Associato presso il Politecnico di Torino, e si occupa di applicazioni della Meccanica Statistica in particolare a problemi di inferenza ed ottimizzazione.

Luca Dall'Asta: è Professore Associato presso il Politecnico di Torino. Si occupa di Fisica Statistica e dei Sistemi Complessi e delle sue applicazioni interdisciplinari.

Alessandro Ingrosso: è *Postdoctoral Fellow* presso l'*Abdus Salam International Centre for Theoretical Physics*. Si occupa di Neuroscienze Computazionali, Machine Learning ed applicazioni interdisciplinari della Meccanica Statistica.

Reti Neurali in grado di apprendere

Una mente intelligente è quella che è in costante apprendimento.

Bruce Lee

Giorgio Buttazzo *Scuola Superiore Sant'Anna, Pisa*

Grazie alle conoscenze acquisite sul cervello umano, l'intelligenza artificiale è riuscita a sviluppare modelli matematici del neurone biologico che oggi vengono utilizzati per costruire reti neurali artificiali, ossia sistemi computazionali in grado di apprendere dai propri errori. Negli anni sono stati sviluppati diversi paradigmi di apprendimento. Quello più noto è il paradigma supervisionato, che consente ad una rete di apprendere delle associazioni attraverso un insieme di esempi preparati da un trainer, che indica alla rete la risposta giusta per ognuno di essi. Un'altra modalità di apprendimento è quella basata su premi e punizioni, che presuppone l'esistenza di un `critico`, il quale, questa volta, non conosce le risposte giuste da suggerire alla rete, ma giudica solo la bontà delle azioni prodotte, penalizzando o incentivando la rete a produrle nuovamente. Infine, nel paradigma non supervisionato, la rete neurale è in grado di auto-organizzarsi in funzione unicamente dei dati che riceve in ingresso, specializzandosi al punto da discriminarli in base alle differenze più salienti rilevate.

L'obiettivo di questo articolo è di introdurre i concetti fondamentali del calcolo neuronale, presentando i vari meccanismi di apprendimento sviluppati negli anni, illustrandone alcuni esempi di utilizzo e le potenzialità future.

Introduzione

La capacità di apprendere dai propri errori e, in generale, di adattarsi ai cambiamenti è una caratteristica essenziale dell'intelligenza. Senza questa capacità, probabilmente la specie umana non sarebbe diventata la specie dominante sul pianeta Terra.

Nel campo dell'intelligenza artificiale, l'importanza dell'apprendimento è stata riconosciuta solo di recente, in quanto le metodologie classiche e gli algoritmi sviluppati tra gli anni sessanta e gli anni 2000, avevano prodotto ottimi risultati in diversi campi applicativi, quali le diagnosi mediche, le previsioni meteorologiche, la dimostrazione automatica di teoremi, la comprensione di testi e i giochi di strategia, come dama e scacchi. Tra i successi più noti, ricordiamo la sfida scacchi-

stica tra Deep Blue (un supercomputer dell'IBM) e il campione del mondo Gary Kasparov, conclusasi l'11 maggio del 1997 con la vittoria di Deep Blue per 3.5 a 2.5.

Fondamentalmente, le tecniche utilizzate in quegli anni dall'intelligenza artificiale si basavano su algoritmi di ricerca su strutture dati ad albero, funzioni euristiche per la valutazione dei risultati, manipolazione e combinazione di regole predefinite. I problemi di tale approccio sono emersi quando i ricercatori hanno cominciato ad utilizzare quegli stessi algoritmi per programmare dei robot a svolgere attività senso-motorie complesse, come la manipolazione di oggetti, la locomozione, il controllo dell'equilibrio, il riconoscimento di immagini e la comprensione del parlato. Gli stessi algoritmi in grado di sconfiggere il campione del mondo di scacchi fallivano miseramente se applicati al riconoscimento di forme o al controllo motorio. Successivamente si è capito che la ragione di tale fallimento è dovuta al fatto che il numero delle possibili situazioni da considerare nello svolgimento di attività senso-motorie è talmente elevato che non è possibile codificare il comportamento di un sistema mediante un insieme di regole predefinite.

Si consideri, ad esempio, di voler sviluppare un programma per il riconoscimento di caratteri manoscritti. La Figura 1 mostra alcuni esempi di immagini (28×28 pixel, ciascuno con 256 livelli di grigio) di caratteri numerici scritti a mano, illustrando la grande differenza che può esistere tra i modi di scrivere lo stesso carattere. Ora immaginiamo di impostare il riconoscimento sulla base di regole del tipo

- **IF** (esiste un tondino in alto a sinistra) **AND** (esiste una linea verticale incurvata verso sinistra) **THEN** (il carattere è il nove).

Affinchè tale regola possa essere compresa da un algoritmo, occorre renderla più precisa, specificando il significato delle frasi "un tondino in alto a sinistra" e "una linea verticale incurvata verso sinistra". Dovrebbe risultare chiaro che, più si cerca di rendere precisa la regola, più sono i casi particolari che occorre considerare per definire il significato delle frasi. Tale approccio è quindi destinato a generare una quantità esorbitante di casi particolari, eccezioni e sottoregole che comunque non coprirebbero tutte le possibilità che si possono presentare. Basti pensare che una piccola matrice binaria di 16×16 pixel, ciascuno con soli due livelli di grigio (bianco e nero) può rappresentare ben $2^{16\times16} \sim 10^{77}$ immagini diverse, ossia un numero paragonabile al numero di atomi presenti nell'intero universo (stimato tra 10^{79} e 10^{81}). Dunque l'approccio a regole è destinato a fallire su questa tipologia di problemi in cui il numero di casi possibili cresce esponenzialmente con la dimensione del dato.

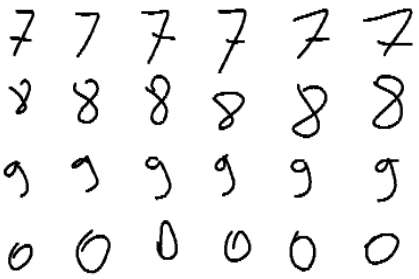


Figura 1: *Esempi di caratteri scritti a mano rappresentati da immagini di 28×28 pixel con 256 livelli di grigio.*

Tuttavia, il cervello umano riesce a risolvere i problemi di riconoscimento e coordinamento senso-motorio in modo rapido ed efficiente. Questa semplice osservazione, unita alle difficoltà di risolvere tali problemi con l'approccio a regole, ha portato i ricercatori a sviluppare dei modelli computazionali ispirati al funzionamento del cervello. Il paragrafo successivo presenta una panoramica storica dei risultati più significativi sulle reti neurali, ottenuti dagli inizi degli anni 40 fino ad oggi.

Evoluzione della ricerca sulle reti neurali

Il neurone binario a soglia

Il primo modello di neurone artificiale, noto come neurone binario a soglia, è stato proposto nel 1943 da due ricercatori statunitensi, `Warren McCulloch` (un neurofisiologo) and `Walter Pitts` (un matematico) [1]. Il modello, schematicamente illustrato in Figura 2, consiste in un elemento di calcolo (neurone artificiale) che riceve n valori di ingresso $(x_1, x_2, \ldots, x_n)$ attraverso altrettanti canali che rappresentano i den-

driti di un neurone biologico. Ciascun valore x_i viene modulato da un peso (*weight*) w_i, che modella la connessione sinaptica presente sul canale dendritico. Gli n valori di ingresso, opportunamente pesati, vengono poi sommati tra loro per produrre il valore di attivazione $a = \sum_{i=1}^{N} w_i x_i$, equivalente al potenziale di membrana di un neurone biologico. Un neurone biologico produce

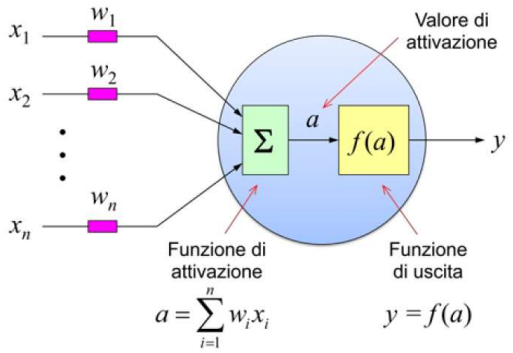


Figura 2: *Modello di neurone binario a soglia proposto da McCulloch e Pitts nel 1943.*

un segnale di uscita (*spike*) quando il potenziale di membrana supera un certo livello di soglia. Analogamente, nel modello di McCulloch e Pitts, il valore di uscita y del neurone viene calcolato come $y = f(a)$, dove $f(.)$ è detta funzione di uscita. Nel neurone binario a soglia, come funzione di uscita si utilizza la funzione di Heaviside, illustrata in Figura 3, corrispondente ad un gradino con soglia θ. L'uscita di un neurone binario a

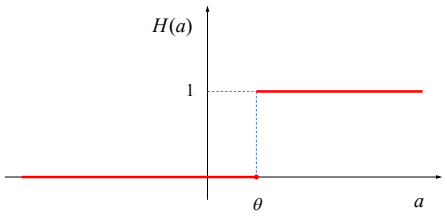


Figura 3: *Funzione di Heaviside, utilizzata come funzione di uscita nel neurone binario a soglia.*

soglia, pertanto, può essere espressa come

$$y = +1 \quad \text{Se} \sum_{i=1}^{N} w_i x_i > \theta,$$
$$y = 0 \quad \text{Altrimenti.}$$

È importante osservare che una differenza sostanziale tra un neurone biologico e il modello binario a soglia è che il neurone biologico codifica l'informazione in frequenza, trasmettendo sull'assone una sequenza di *spike* con frequenza proporzionale ai segnali ricevuti in ingresso, mentre il neurone binario a soglia codifica l'informazione in ampiezza, quantizzata su due valori di uscita (0) e (1). Un'altra differenza importante è che il modello di McCullogh e Pitts non è in grado di apprendere, anche perchè in quegli anni non erano stati ancora compresi i meccanismi dell'apprendimento.

La scoperta di Hebb

Nel 1949, lo psicologo canadese `Donald Hebb` [2] fece una scoperta rivoluzionaria che avanzò le conoscenze sul cervello e fece progredire la ricerca sulle reti neurali. Hebb scoprì che il processo di apprendimento non modifica di fatto il funzionamento delle cellule nervose, ma opera unicamente sulle connessioni sinaptiche che modulano la comunicazione tra i neuroni. Egli riportò che

> "Quando un assone di una cellula A è abbastanza vicino da eccitare una cellula B e partecipa ripetutamente alla sua attivazione, si osservano alcuni processi di crescita o cambiamenti metabolici in una o entrambe le cellule tali da aumentare l'efficacia di A nell'attivare B."

Questa nuova conoscenza sui meccanismi dell'apprendimento ha consentito di integrare questa capacità anche nei modelli neurali allora sviluppati.

Il Perceptron

Grazie alla scoperta di Hebb, nel 1957, lo psicologo statunitense `Frank Rosenblatt` [3] sviluppò il primo modello di neurone artificiale in grado di apprendere, il **Perceptron**, illustrato in Figura 4. La regola di apprendimento, nota come **Delta Rule** [4], è semplice ma efficace e consiste nel modificare i pesi in modo proporzionale all'ingresso e all'errore (δ) commesso dal neurone, pari alla differenza tra l'uscita reale (y) e quella desiderata (y_d). La costante di proporzionalità

è detta *learning rate*. In un esperimento che di-

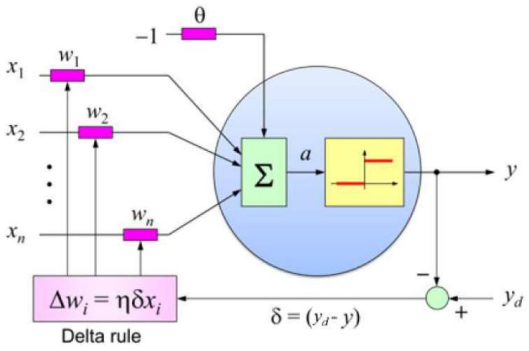


Figura 4: *Il Perceptron di Rosenblatt. Durante la fase di apprendimento, i pesi vengono modificati in funzione dell'errore* (δ) *commesso dal neurone, pari alla differenza tra l'uscita desiderata* (y_d) *e quella reale* (y).

venne poi famoso, Rosenblatt costruì il primo `Perceptron` in *hardware*, realizzando i pesi con dei potenziometri motorizzati e collegando gli ingressi a 400 fotocellule disposte come una matrice di 20×20 *pixel*. Presentando in ingresso i disegni di varie forme geometriche, Rosenblatt riuscì ad addestrare il `Perceptron` a riconoscere le forme concave da quelle convesse[1].

Il controesempio di Minsky e Papert

Gli entusiasmi di Rosenblatt purtroppo svanirono nel 1969, quando due matematici del Massachusetts Institute of Technology (MIT), `Marvin Minsky` e `Seymour Papert`, pubblicarono un libro intitolato "Perceptrons" [5], in cui venivano dimostrati formalmente i punti di forza ma anche i maggiori limiti del modello di Rosenblatt. In particolare, Minsky e Papert dimostrarono, attraverso un controesempio, l'impossibilità per un `Perceptron` di apprendere la semplice funzionalità di un OR esclusivo (XOR) a due ingressi, che prevede una risposta pari a zero quando i due ingressi sono uguali (entrambi zero o entrambi uno) e una risposta pari a uno quando i due ingressi sono diversi.

Questo risultato negativo sul Perceptron fece precipitare l'interesse per le reti neurali per oltre un decennio, dal 1969 al 1982, periodo che oggi viene indicato come *AI winter*.

Le reti di Hopfield

L'interesse per le reti neurali si riaccese nel 1982, quando `John Hopfield` [6] propose un nuovo modello di rete in grado di comportarsi come una memoria associativa, ossia una memoria in cui è possibile memorizzare un insieme di informazioni desiderate per poi recuperarle partendo da dati parziali o distorti. Se ad esempio in una rete di Hopfield vengono memorizzate delle immagini, queste possono poi essere recuperate fornendo in ingresso alla rete delle immagini simili, rumorose o distorte.

Hopfield dimostrò che tale proprietà può essere ottenuta costruendo una rete di neuroni binari a soglia che soddisfi le seguenti proprietà:

1. tutti i neuroni sono connessi tra loro;
2. la funzione di uscita è la funzione segno;
3. ogni coppia di neuroni ha pesi simmetrici;
4. i neuroni cambiano stato uno per volta.

Se queste proprietà vengono rispettate, è possibile dimostrare che, partendo da un qualsiasi stato iniziale, la rete evolve attraverso una serie di commutazioni, generando una sequenza di stati (traiettoria) che termina sempre in uno stato stabile.

Hopfield fornì anche una regola per poter rendere stabili degli stati neurali desiderati, che rappresentano quindi le memorie della rete. Per rendere stabile una configurazione di attivazioni neurali, egli suggerì di collegare neuroni con attivazione simile con pesi positivi e neuroni con attivazione opposta con pesi negativi. Per memorizzare più informazioni stabili basterà sommare i pesi ottenuti per ciascuno stato.

La Figura 5 illustra un esempio di come la rete di Hopfield sia in grado di recuperare l'immagine del numero tre (precedentemente memorizzata come stato stabile) partendo da un'immagine notevolmente rumorosa presentata in ingresso.

Le reti di Kohonen

Sempre nel 1982, `Teuvo Kohonen` propose un modello di rete neurale [7] capace di autoorganizzarsi per formare delle mappe sensoriali

[1] Si ricorda che una forma si dice convessa se, presi due punti A e B al suo interno, il segmento che li unisce è contenuto tutto all'interno della figura. Viceversa la figura si dice concava.

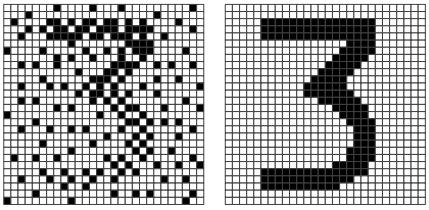


Figura 5: *Esempio di memoria associativa realizzata mediante una rete di Hopfield: (a) immagine presentata come stato iniziale; (b) immagine recuperata, corrispondente ad uno stato stabile memorizzato in precedenza. Ogni pixel corrisponde ad un neurone. In questo esempio la rete ha 784 neuroni per rappresentare immagini binarie di* 28×28 *pixel.*

simili a quelle esistenti nella corteccia somatosensoriale, sulla quale viene rappresentato il cosiddetto homunculus sensitivo. Una rete neurale di Kohonen è formata da due soli strati: uno stato di ingresso e uno di uscita, come rappresentato in Figura 6. La regola di apprendimento è tale

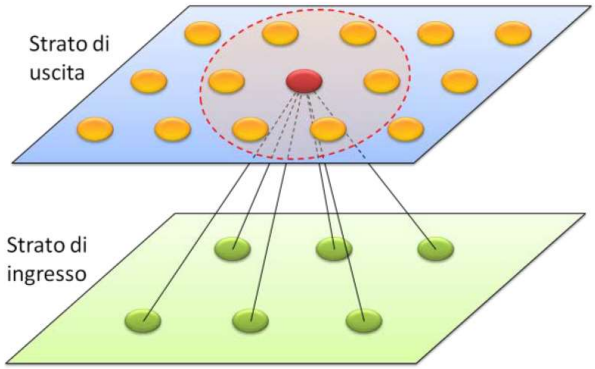


Figura 6: *Esempio di una rete di Kohonen con 6 neuroni di ingresso e 15 neuroni di uscita disposti su una mappa bidimensionale.*

da creare un isomorfismo tra stimoli sensoriali di ingresso e neuroni di uscita, per cui neuroni vicini si specializzano a riconoscere stimoli sensoriali simili. Questa proprietà viene ottenuta attraverso un meccanismo di apprendimento competitivo che, per ogni stimolo, aggiudica come vincitore il neurone che ha l'attivazione più alta tra tutti. Per ottenere l'isomorfismo, i pesi del neurone vincitore e quelli dei neuroni appartenenti ad un vicinato (illustrato in figura con una linea tratteggiata rossa) vengono modificati in modo che tali neuroni si specializzino ancor meglio a riconoscere quello stimolo.

Le reti di Kohonen assumono una grande rilevanza nel panorama dei modelli neurali, in quanto consentono l'estrazione di caratteristiche salienti dai dati di ingresso senza alcuna supervisione da parte dell'utente. Esse vengono utilizzate per ottenere una compressione dei dati, o un raggruppamento di dati omogenei in un insieme di classi (*clustering*) in base alla somiglianza tra i dati. Esse possono persino essere utilizzate per risolvere efficientemente problemi di ottimizzazione combinatoria.

Reinforcement Learning

Nel 1983, Andrew Barto, Richard Sutton e Charles Anderson [8] proposero un nuovo modello di rete neurale in grado di generare azioni di controllo utilizzando un paradigma di apprendimento basato su premi e punizioni, e denominato Reinforcement Learning.

L'idea alla base di questo meccanismo è che la rete neurale generi inizialmente delle azioni casuali di controllo e riceva una ricompensa (un segnale di *feedback* positivo) o una punizione (un segnale di *feedback* negativo) in base all'esito di tali azioni. I segnali di *feedback* ricevuti vengono utilizzati per modificare i pesi della rete in modo da favorire le azioni che hanno generato una ricompensa e scoraggiare quelle che hanno generato una punizione. In questo modo la rete si costruisce gradualmente una conoscenza del sistema, passando da una fase esplorativa, pesantemente guidata dal caso, ad una fase operativa, in cui la conoscenza acquisita viene sfruttata per generare le azioni migliori.

Dal 1983 ad oggi, questo paradigma di apprendimento si è notevolmente evoluto, grazie anche alle tecniche più avanzate di *deep learning* sviluppate di recente, raggiungendo prestazioni eccellenti in diverse applicazioni. In particolare, nel 2010, la DeepMind Technology ha utilizzato questa metodologia per addestrare una rete neurale a giocare a numerosi videogiochi Atari, riuscendo a superare le prestazioni umane in ben sette di essi. Nel 2014, l'azienda è stata acquisita da Google e il 23 maggio 2017 la Google DeepMind è ritornata alla ribalta per aver costruito una rete basata su *reinforcement learning*, denominata

AlphaGo Zero, in grado di battere il campione del mondo di Go, Ke Jie. Questo risultato è alquanto rilevante, poichè il Go è uno tra i giochi più complessi al mondo, in cui l'albero delle possibili posizioni è dell'ordine di 10^{170}, contro il 10^{120} degli scacchi.

Oggi, il *reinforcement learning* è tra gli algoritmi maggiormente studiati, in quanto promette di risolvere una grande quantità di problemi rilevanti e difficili, tra cui la l'apprendimento della camminata in robot bipedi, la guida di veicoli autonomi e il controllo di sonde esplorative spaziali, solo per citarne alcuni.

La Backpropagation

Nel 1986, David Rumelhart, Geoffrey Hinton e Ronald Williams [9] svilupparono un potente algoritmo di apprendimento supervisionato, noto come Backpropagation, che permette ad una rete neurale di imparare a classificare dei pattern di ingresso attraverso un insieme di esempi, detto *training set*.

Le reti neurali addestrabili con Backpropagation sono di tipo stratificato, come quella illustrata in Figura 7. Ogni neurone di uno strato è connesso con ogni neurone dello strato successivo, ma non esistono connessioni tra neuroni dello stesso strato, nè tra neuroni appartenenti a strati non adiacenti. Il primo strato è quello di ingresso *(input layer)*, che riceve i dati da elaborare. L'ultimo strato è quello di uscita *(output layer)*, che produce i risultati dell'elaborazione. Gli strati intermedi vengono detti strati nascosti *(hidden layer)* in quanto non sono visibili dall'esterno in una visione *black-box* della rete. In questo tipo di rete, il modello di neurone utilizzato in tutti gli strati è molto simile al neurone binario a soglia e differisce unicamente per la funzione di uscita. Le funzioni di uscita oggi più utilizzate sono la sigmoide, la tangente iperbolica e la lineare rectificata (ReLU), illustrate in Figura 8. L'aspetto più interessante dell'apprendimento supervisionato è che la rete riesce a generalizzare ciò che ha appreso, classificando correttamente nuovi dati mai visti in fase di *training*.

Grazie a questi risultati, nei vent'anni successivi alla nascita della Backpropagation, le reti neurali sono state utilizzate per risolvere diverse

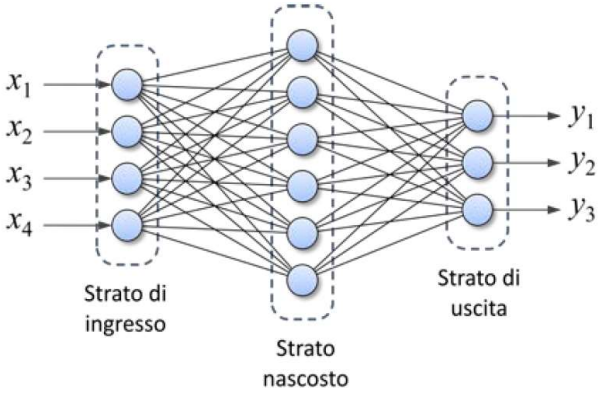


Figura 7: *Esempio di rete a tre strati addestrabile con l'algoritmo di Backpropagation.*

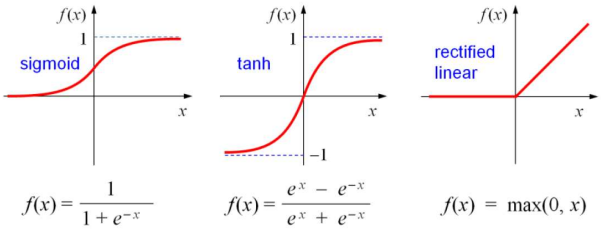


Figura 8: *Funzioni di uscita comunemente utilizzate nelle reti multistrato addestrate con Backpropagation.*

tipologie di problemi, tra cui il riconoscimento di immagini, la compressione di dati, la previsione di segnali e serie storiche e il controllo di sistemi robotici, nei più disparati settori, quali fisica, chimica, ingegneria, geologia, agraria, astronomia, economia, medicina, scienze sociali, psicologia, ecc.

Nuovi ostacoli

Nonostante l'esplosione dei campi applicativi, dal 1986 al 2006 non ci furono nuovi sostanziali sviluppi teorici sulle reti neurali. Molti ricercatori provarono a sviluppare modelli neurali più complessi, più vicini alla controparte biologica, ma non si riusciva ad ottenere dei comportamenti interessanti e, soprattutto, a dimostrare delle proprietà generali sulle le quali costruire algoritmi generalizzabili. Altri ricercatori provarono ad aumentare il numero di strati di una rete addestrata con l'algoritmo di Backpropagation, ma osservarono grosse difficoltà ad addestrare reti con più di quattro strati. Tali problemi sono stati risolti solo negli anni 2000.

La rivoluzione delle deep neural network

A partire dal 2006, la ricerca sulle reti neurali ha avuto una grossa impennata grazie e tre importanti fattori:

1. Compresi i problemi che causavano la difficoltà di addestrare reti con più di quattro strati, sono state sviluppate nuove metodologie in grado di superare quei limiti e gestire l'apprendimento di reti molto più grandi, costituite da migliaia di neuroni organizzati su numerosi strati: le `deep neural network`.

2. Le `deep neural network`, essendo costituite da migliaia di neuroni, richiedono una grossa potenza di calcolo per essere addestrate. Intorno al 2006, la potenza di calcolo necessaria è diventata disponibile a basso costo grazie alla produzione di nuove piattaforme di calcolo basate sulle `Graphics Processing Unit` (GPU). Tali piattaforme, originariamente progettate per gestire operazioni grafiche, sono state modificate per poter svolgere anche calcoli vettoriali, come quelli necessari in una rete neurale.

3. I primi risultati ottenuti con le `deep neural network` hanno attratto l'interesse di grosse aziende, come `Google`, `Microsoft` e `Facebook` che, gestendo un'enorme quantità di dati, hanno visto nelle reti neurali una grossa opportunità per risolvere problemi di classificazione di immagini, riconoscimento di volti, suoni, voci, e hanno quindi cominciato ad investire grosse quantità di risorse in questo settore.

Infine, un altro elemento che ha contribuito all'evoluzione delle `deep network` è stata la competizione internazionale `ImageNet`, o più precisamente la `ImageNet Large-Scale Visual Recognition Challenge` (ILSVRC) [10], una sorta di olimpiade annuale della *computer vision*, nata nel 2010 per mettere in competizione i migliori gruppi al mondo su problemi complessi relativi al riconoscimento di immagini. Ogni squadra aveva a diposizione un enorme database e doveva addestrare la propria rete su un milione di immagini suddivise in mille categorie. Una volta addestrate, le reti dovevano classificare 100.000 nuove immagini. Poichè un'immagine poteva contenere diversi oggetti, la risposta della rete era considerata corretta se la classe veniva correttamente identificata considerando le 5 uscite con valore più elevato (sulle mille possibili). La Figura 9 illustra come si è ridotto l'errore di classificazione delle reti vincitrici dal 2010 al 2017.

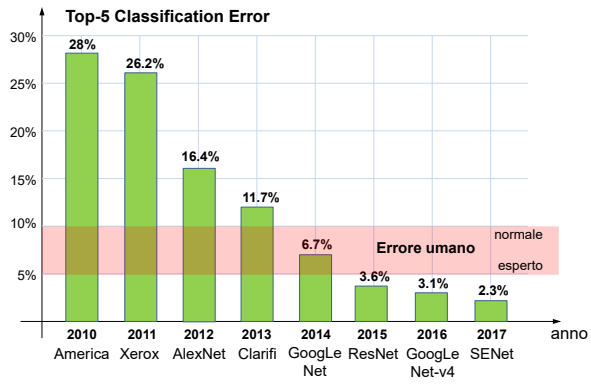


Figura 9: *Diminuzione dell'errore di classificazione delle reti neurali dal 2010 al 2017 nella competizione ImageNet.*

È interessante notare come, per la prima volta nella storia, nel 2014 la rete `GoogLeNet` ha eguagliato le prestazioni umane, poi superate negli anni successivi da altre reti.

Paradigmi di Apprendimento

Negli anni sono stati proposti diverse modalità di apprendimento, che possono essere ricondotte e tre paradigmi principali:

1. apprendimento con supervisione;

2. apprendimento senza supervisione;

3. apprendimento con rinforzo.

I concetti alla base di questi meccanismi sono descritti di seguito.

Apprendimento supervisionato

L'obbiettivo dell'apprendimento supervisionato è quello di addestrare una rete neurale a riconoscere dei dati di ingresso come appartenenti a delle `categorie (o classi)` predefinite. Tali categorie vengono mostrate alla rete mediante un insieme di esempi, detto *training set*.

Il funzionamento della rete è suddiviso in due fasi, dette `di addestramento` e `di inferenza`. Nella fase di addestramento vengono presentati gli esempi del training set dai quali la rete deve imparare. Ciascun esempio consiste in una coppia di dati vettoriali: l'ingresso da classificare (x) e l'uscita desiderata da associare (y_d). Per ogni esempio viene calcolata l'uscita della rete (y) e tali valori sono utilizzati per calcolare una `funzione di errore` (detta anche `loss function`) con cui modificare i pesi della rete.

Detta E la funzione che descrive l'errore della rete rispetto all'uscita desiderata y_d, ciascun peso viene modificato in modo da diminuire l'errore, calcolando il gradiente della funzione E rispetto al peso. Uno schema del paradigma supervisionato è illustrato in Figura 10. Ad ogni iterazione l'errore tende a diminuire e, quando esso diventa inferiore ad una certa soglia prestabilita, la rete si considera addestrata e può essere utilizzata per la fase di inferenza. In questa fase, la rete viene utilizzata per effettuare la classificazione vera e propria di nuovi dati. Al fine di comprendere

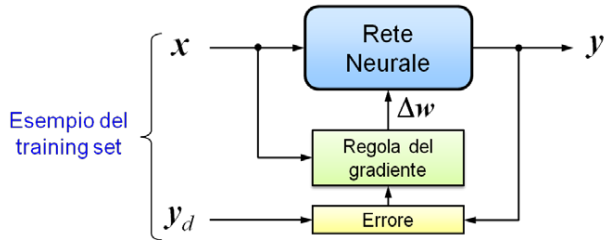


Figura 10: *Paradigma di apprendimento supervisionato.*

meglio il paradigma supervisionato, si consideri una rete con due ingressi (x_1, x_2) e una sola uscita (y). Tale rete può essere addestrata a riconoscere se un certo insieme di dati appartiene ad una certa classe A $(y_d = 1)$ oppure no $(y_d = 0)$. Visto che ogni dato d'ingresso è descritto da una coppia di coordinate, gli esempi possono essere rappresentati su un piano cartesiano con dei punti: se un dato appartiene alla classe A allora il punto viene indicato con il colore rosso, altrimenti con il colore nero. La situazione descritta è illustrata in Figura 11.

Terminato l'apprendimento, nella fase di inferenza la rete deve stabilire se dei nuovi dati (rappresentati in figura con dei punti grigi) appartengano o meno alla classe A.

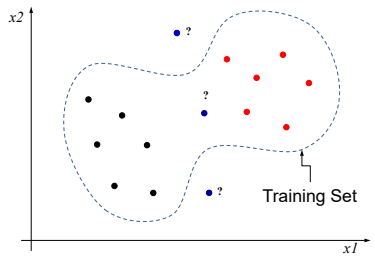


Figura 11: *Problema della classificazione nel paradigma di apprendimento supervisionato. I punti colorati rappresentano gli esempi del training set descritti da due variabili (x_1, x_2): i punti rossi appartengono ad una classe A, i punti neri no. Terminato l'apprendimento, nella fase di inferenza la rete deve stabilire se dei nuovi dati (i punti blu) appartengono o meno alla classe A.*

Apprendimento senza supervisione

Nell'apprendimento senza supervisione gli esempi utilizzati per addestrare la rete non sono etichettati come appartenenti ad una classe, per cui non esiste un'uscita desiderata per ogni dato di ingresso. In questo caso, la rete deve imparare a suddividere i dati in diversi gruppi (*cluster*) sulla base della loro somiglianza. La Figura 12 illustra un caso in cui i 12 esempi forniti alla rete vengo raggruppati in due *cluster* di 6 elementi ciascuno. La figura mostra anche come,

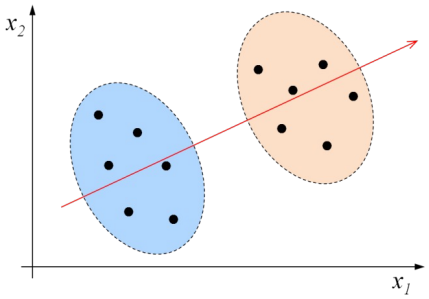


Figura 12: *Problema del* clustering *nel paradigma di apprendimento non supervisionato. In questo caso i dati non sono etichettati come appartenenti ad una classe. Lo scopo della rete è di suddividerli in gruppi (cluster) sulla base della loro somiglianza.*

in molti casi, la rete può rilevare una ridondanza nei dati e operare una riduzione di dimensioni. Nel caso di figura, ad esempio, i *cluster* possono essere separati utilizzando una sola variabile

(rappresentata dall'asse indicato in rosso) ottenuta come combinazione lineare delle due variabili originali. Questa proprietà, comune a molte reti non supervisionate, rende questo meccanismo di apprendimento particolarmente adatto alla compressione di dati oppure all'estrazione di caratteristiche salienti dai dati di ingresso.

Apprendimento con Rinforzo

Il paradigma di apprendimento con rinforzo si utilizza tipicamente nei problemi di controllo, ossia quando vogliamo addestrare una rete neurale ad inviare delle azioni di comando ad un sistema che interagisce con un ambiente. Tale modalità di apprendimento è una via di mezzo fra le due precedenti, poichè richiede solo una leggera supervisione, che però non necessita di fornire la risposta desiderata per ognuno degli esempi del training set. Per ogni azione generata sul sistema, la rete riceve una valutazione da parte di un `critico`, la cui funzione è solo quella di accorgersi quando il sistema fallisce oppure raggiunge un obiettivo. Tale valutazione è codificata in un `segnale di rinforzo` o *reward* che viene utilizzato per modificare i pesi della rete.

Per fare un esempio concreto, si consideri lo schema illustrato in Figura 13, supponendo di utilizzare l'uscita della rete neurale per controllare lo sterzo di un auto. Al fine di poter imparare a sterzare correttamente, la rete dovrà ricevere delle informazioni sullo stato dell'auto, ad esempio le immagini prelevate da una telecamera che inquadra la strada. In questo caso, il segnale di rinforzo (R) prodotto dal critico potrebbe essere negativo (-1) quando l'auto esce fuori strada e positivo (1) quando l'auto riesce a tenersi al centro della carreggiata. In tutti gli altri casi, il segnale di rinforzo può essere nullo. L'obiettivo dell'apprendimento con rinforzo è quindi quello di imparare a generare azioni che migliorino la valutazione del critico nel tempo. L'esempio illustrato suggerisce come questa modalità di apprendimento sia paragonabile a quella basata su premi e punizioni. Un rinforzo positivo ricevuto dal critico è assimilabile ad un premio, mentre un rinforzo negativo è assimilabile ad una punizione. Il meccanismo di apprendimento è tale da scoraggiare la rete a ripetere le azioni che in certo stato hanno generato fallimenti, favorendo invece le azioni che hanno causato delle valutazioni positive.

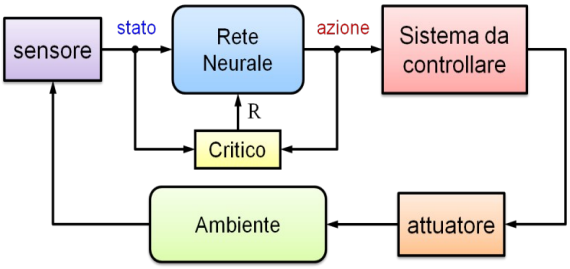


Figura 13: *Paradigma di apprendimento con rinforzo. La rete neurale genera delle azioni di controllo su un sistema che interagisce con l'ambiente e riceve da un critico una valutazione (R) sulla loro efficacia. Lo scopo dell'apprendimento con rinforzo è di imparare a generare azioni che migliorino la valutazione del critico.*

Considerato che il critico può essere facilmente realizzato elaborando i dati prodotti da opportuni sensori (ad esempio sensori di contatto, accelerazione, distanza, ecc.) questo paradigma di apprendimento risulta molto potente, poichè in grado di scoprire le azioni corrette senza l'intervento umano, ma unicamente sulla base dei fallimenti e dei successi sperimentati dal sistema.

I modelli recenti che hanno rivoluzionato l'intelligenza artificiale

Dal 2000 ad oggi sono stati ideati nuovi modelli di rete neurale che hanno permesso di risolvere problemi prima considerati intrattabili con queste tecniche. In ordine temporale, i modelli più rilevanti proposti in letteratura sono le reti ricorrenti, le reti convoluzionali e le reti generative.

Le Reti Ricorrenti

A differenza delle reti neurali considerate finora in questo articolo, le `reti neurali ricorrenti`, o *Recurrent Neural Networks* (RNN) sono in grado di trattare sequenze temporali di dati sia in

ingresso che in uscita. Per ottenere questo scopo, le reti ricorrenti posseggono delle connessioni aggiuntive, rispetto alle reti multistrato, che collegano le uscite dei neuroni nascosti ai neuroni di ingresso. La Figura 14 illustra la struttura di una rete ricorrente, in cui l'uscita dello strato nascosto $\mathbf{h}(t)$ viene riportata in ingresso (a). Una rete ricorrente viene spesso rappresentata in una modalità sviluppata nel tempo (*unrolled*), in cui sono evidenziati i valori dei vettori per i vari istanti temporali (b). Tra i moti convoluzionali, strati di subsampling e strati completamente connessi.

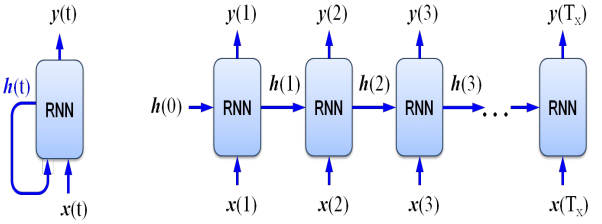


Figura 14: *Struttura di una rete ricorrente in cui l'uscita dello strato nascosto* $\mathbf{h}(t)$ *viene riportata in ingresso* (a)*. La figura* (b) *illustra la versione* unrolled *in cui sono evidenziati i valori dei vettori per i vari istanti temporali.*

delli più utilizzati per le reti ricorrenti ricordiamo le `Long short-term memory` (LSTM) [11] e le `Gated Recurrent Units` (GRU) [12]. Tra le applicazioni più interessanti di questi modelli citiamo l'analisi e la previsione di testi, il riconoscimento vocale, la traduzione automatica, il riconoscimento di sequenze video, la descrizione testuale di immagini e la composizione musicale.

Le Reti Convoluzionali

Le reti convoluzionali sono state introdotte nel 1998 da `Yann LeCun` et al. [13], sebbene delle versioni preliminari fossero state già utilizzate nel 1970. Tali reti sono particolarmente indicate per la classificazione di immagini, in quanto fanno uso di una speciale architettura che sfrutta la struttura spaziale delle immagini per ridurre il numero di connessioni tra neuroni. La tipica architettura di una rete convoluzionale consiste in una sequenza di strati di vario tipo, tra cui strati convoluzionali, strati di subsampling e strati completamente connessi.

Gli strati convoluzionali effettuano un'estrazione di caratteristiche dallo strato precedente attraverso un'operazione di convoluzione, che consiste nel moltiplicare i valori x di una piccola area di dimensione $r \times r$ (*receptive field*) per una matrice w di pesi (*weights*) della stessa dimensione, detta *kernel* o `filtro`. Tale filtro viene traslato sullo strato in modo da estrarre la stessa caratteristica in zone diverse dello strato. In questo modo, il numero di pesi da modificare risulta pari ad r^2 ed è indipendente dalle dimensioni dello strato.

La Figura 15 illustra un esempio di convoluzione su uno strato di 129×129 neuroni, utilizzando un filtro di dimensioni 3×3 traslato di 2 neuroni per volta. Lo strato convoluzionale risulta pertanto composto da 64×64 neuroni, che costituiscono una mappa di caratteristiche (*feature map*) rilevate in diverse posizioni spaziali. Ciascun neurone dello strato convoluzionale risulta quindi un rivelatore della caratteristica codificata nei pesi nel filtro: un valore elevato indica che quella caratterista è presente nel campo recettivo corrispondente. Gli strati di *subsampling* effettuano

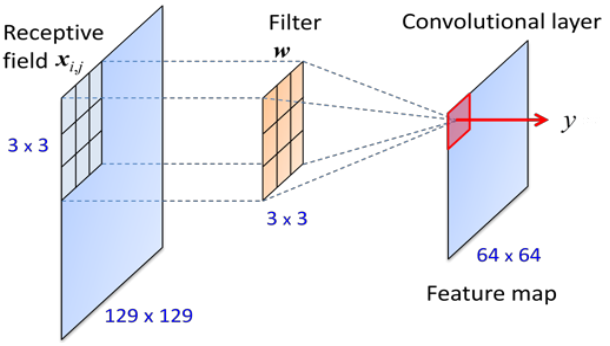


Figura 15: *Operazione di convoluzione elementare in una rete convoluzionale. Il valore* y *del neurone in rosso è calcolato come la somma dei prodotti tra i valori dei neuroni del campo recettivo* x *e i pesi del filtro* w*, i.e.* $y = \mathbf{x_{i,j}} * \mathbf{w} = \sum_{l=1}^{r} \sum_{m=1}^{r} x_{i+l,j+m} w_{lm}$.

una compressione dell'informazione, riducendo il numero di neuroni dello strato precedente attraverso operazioni di media o di massimo su piccole aree neuronali. Una rete di solito utilizza diversi strati convoluzionali seguiti da altrettanti strati di *subsampling*. Gli strati completamente connessi realizzano infine la classificazione vera

e propria e sono utilizzati nella parte finale della rete. La rete termina con lo strato di uscita, il cui numero di neuroni è pari al numero di categorie che si desidera riconoscere nelle immagini.

Un'altra funzione importante realizzabile attraverso le reti convoluzionali è quella di rilevare anche la posizione dell'oggetto riconosciuto (*object detection*) attraverso un rettangolo (*bounding box*) descritto da quattro coordinate, due per il centro e due per le dimensioni dei lati. In questa modalità operativa, se C è il numero di categorie da riconoscere, lo strato di uscita dovrà contenere almeno $C + 4$ neuroni.

Oggi le reti convoluzionali hanno raggiunto prestazioni eccellenti in diversi settori applicativi, tra cui quello medico, in cui le reti neurali sono state utilizzate per effettuare diagnosi precoci a partire da immagini mediche e scansioni tomografiche. In particolare sono state ottenute prestazioni paragonabili o superiori a quelle umane nell'analisi di elettrocardiogrammi, nell'identificazione di tumori della pelle e di patologie della retina, del cancro al polmone e nella diagnosi dell'Alzheimer.

Le Reti Generative

Le reti generative o *Generative Adversarial Networks* (GAN) sono una particolare classe di reti neurali ideate nel 2014 da `Ian Goodfellow` et al. [14] al fine di produrre nuovi campioni di dati aventi la stessa distribuzione statistica di quelli utilizzati per l'addestramento. In altre parole, se una GAN viene addestrata utilizzando un database di volti umani, alla fine dell'addestramento essa sarà in grado generare delle nuove immagini realistiche di volti umani; se addestrata con foto di paesaggi naturali, essa potrà generare foto di nuovi paesaggi.

Ma l'utilizzo delle GAN non si riduce a questo. Negli ultimi anni esse sono state utilizzate nelle più svariate applicazioni, tra cui la colorazione di immagini e filmati in bianco e nero (*colorization*), la generazione di immagini a risoluzione più elevata di quella del campione in ingresso (*super resolution*), il restauro di foto danneggiate (*pixel restoration*), la generazione di voci e musica, l'animazione di volti dipinti (*reenactment*), o la trasformazione di foto in quadri stile Van Gogh o Monet (*style transfer*).

Una rete GAN è composta in realtà da due tipi di reti neurali, un `Generatore` ed un `Discriminatore`, che competono tra loro in una sorta di gioco. Il `Generatore` gioca il ruolo di un falsario che produce delle opere false che vuole far passare come autentiche, mentre il `Discriminatore` agisce come un ispettore che cerca di identificare le opere false. La regola di apprendimento è strutturata in modo che entrambe le reti siano portate a migliorare le loro capacità, fino al punto che le opere false diventano indistinguibili da quelle autentiche.

Il `Discriminatore` è addestrato in modo supervisionato per distinguere i campioni veri dai falsi. Esso ha quindi un solo neurone di uscita, che vale 1 quando il campione d'ingresso è autentico e 0 quando è falso. Il `Generatore` è invece addestrato con una modalità non supervisionata, modificando i pesi in modo da favorire la generazione di quei campioni che hanno ingannato il `Discriminatore` e penalizzare la generazione di quelli che sono stati intercettati come falsi. Una volta che la GAN è stata addestrata, il `Discriminatore` viene eliminato, in quanto lo scopo finale è quello di generare campioni realistici. Lo schema architetturale di una GAN è illustrato sinteticamente in Figura 16.

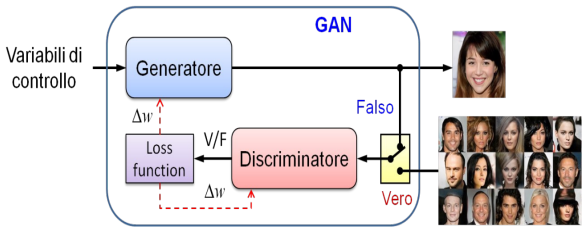


Figura 16: *Architettura di una rete GAN. Il Generatore ha il compito di produrre campioni falsi realistici, mentre il Discriminatore ha il compito di distinguere i campioni autentici (provenienti dal database) da quelli falsi prodotti dal Generatore. Gli errori commessi da ciascuna rete contribuiscono al miglioramento di entrambe.*

Problemi irrisolti

Nei paragrafi precedenti sono state descritte le capacità di elaborazione delle reti neurali in nu-

merosi campi applicativi. Grazie a queste potenzialità, l'industria sta considerando seriamente di utilizzare questa metodologia per sviluppare robot intelligenti e veicoli a guida autonoma. Tuttavia, quando si tratta di realizzare sistemi che devono interagire con l'uomo, occorre garantire non solo prestazioni elevate, ma soprattutto altre proprietà particolarmente critiche, quali predicibilità, affidabilità, e sicurezza.

Fino ad oggi le reti neurali e la maggior parte degli algoritmi di intelligenza artificiale sono stati utilizzati per realizzare applicazioni non critiche, come il riconoscimento di caratteri manoscritti, il riconoscimento facciale, la traduzione automatica, e il riconoscimento vocale. Un errore commesso da una rete in una di queste applicazioni non comporta danni per l'uomo. Immaginiamo invece cosa potrebbe accadere se una rete dovesse commettere un errore di riconoscimento o di controllo in un'automobile a guida autonoma. Del resto, gli incidenti che si sono verificati di recente su automobili avanzate, come la Tesla, che consentono di attivare funzionalità automatiche (sotto la stretta supervisione del conducente), indicano che questi sistemi possono fallire in casi particolari (indicati come *corner case*) in cui più condizioni non previste contribuiscono a produrre un malfunzionamento.

Si consideri, ad esempio, l'incidente verificatosi il 7 Maggio 2016, in cui Joshua Brown è stato vittima di un incidente mortale mentre era alla guida della sua `Tesla S` a guida autonoma nei pressi di Williston, Florida (USA). In base alla ricostruzione dell'incidente [15] si è capito che un TIR che viaggiava sulla corsia opposta ha attraversato la doppia linea gialla per svoltare a sinistra. Il sistema di visione non lo ha rilevato poichè il cielo era troppo luminoso e il camion era bianco, per cui l'auto non ha frenato e si è scontrata contro il TIR, causando la morte di Brown. In base alle direttive `Tesla`, il Sig. Brown sarebbe dovuto intervenire sui comandi, ma purtroppo non lo ha fatto in quanto era intento a guardare un film.

Al fine di evitare questo tipo di incidenti, le auto del futuro dovranno essere progettate per essere tolleranti ai malfunzionamenti di un singolo componente, prevedendo una ridondanza di sottosistemi basati su tecnologie differenti.

Un altro aspetto essenziale da garantire nei sistemi critici è quello della sicurezza. Nel 2015, due ricercatori di *cyber-security*, Charlie Miller and Chris Valasek, sono riusciti da remoto a compromettere il *software* di controllo di una Jeep Cherokee [17], portandola fuori strada dopo aver assunto il controllo del volante e disabilitato la trasmissione e i freni della vettura. Gli strumenti *software* con i quali oggi vengono gestite le reti neurali non sono progettati per essere sicuri, anzi contribuiscono ad aumentare notevolmente le superfici di attacco al sistema di controllo. Pertanto, occorre investire in questo settore di ricerca per mettere in sicurezza tutti i sistemi basati su intelligenza artificiale e ridurre le probabilità di attacco.

Infine, un altro aspetto relativo alla sicurezza delle reti neurali è legato ad una tecnica in grado di generare immagini (dette *adversarial sample*) [16, 18] che appaiono normali alla vista umana, ma vengono interpretate erroneamente dalla rete neurale, che riconosce l'immagine come appartenente ad una categoria diversa, definita arbitrariamente. Sfruttando questa tecnica, un *hacker* potrebbe decidere di attaccare un veicolo autonomo basato su reti neurali senza intervenire sul *software* di controllo, ma semplicemente agendo sull'ambiente. Basterebbe generare un'immagine avversaria di un segnale stradale, ad esempio lo stop, e attaccarla sul vero segnale, in modo che esso venga interpretato come un albero o un uccello.

Molte delle tecniche utilizzate per generare immagini avversarie si basano sulla conoscenza della rete e dei pesi. Esse sfruttano l'algoritmo di `backpropagation` non per modificare il valore dei pesi della rete, ma per modificare il valore di alcuni `pixel` di un'immagine di rifermento posta in ingresso. I `pixel` vengono modificati per ridurre l'errore su una classe erronea desiderata e aumentarlo sulla quella corretta. L'immagine così modificata risulta pressochè identica a quella di partenza ad un osservatore umano, ma la rete neurale la interpreta in modo totalmente diverso.

I ricercatori hanno già cominciato a studiare nuove tecniche per affrontare questo nuovo tipo di attacco, ma non esiste ancora una soluzione definitiva al problema.

Conclusioni

In questo articolo è stata presentata una panoramica della ricerca sulle reti neurali, dai primi modelli sviluppati verso la metà del secolo scorso fino alle metodologie più recenti relative alle deep network. In numerosi settori applicativi le reti neurali hanno raggiunto o superato le prestazioni umane e promettono di diventare uno strumento essenziale per la previsione di eventi, le diagnosi mediche, la progettazione di nuovi farmaci, la guida autonoma e la percezione artificiale nei robot del futuro.

Tuttavia, al fine di poter essere utilizzate in sistemi ad elevata criticità, in cui è prevista una stretta interazione con l'uomo, è necessario affrontare nuovi problemi, finora trascurati, quali la predicibilità temporale delle risposte, l'affidabilità di funzionamento e gli aspetti legati alla sicurezza, al momento ancora irrisolti.

Superate queste difficoltà, in un futuro non tanto lontano, dovremo affrontare un problema ancora più grande, che coinvolgerà aspetti legali, sociali, etici e psicologici: la convivenza con entità artificiali, fisiche e virtuali, dotate di intelligenza superiore a quella umana.

[1] W. S. McCulloch and W. Pitts: "A logical calculus of the ideas immanent in nervous activity", *Bulletin of Mathematical Biophysics* **5** (1943) 115.

[2] D. O. Hebb: *The organization of behavior.* Springer, Berlin (1949).

[3] F. Rosenblatt: *The Perceptron – a perceiving and recognizing automaton*, Report 85–460–1, Cornell Aeronautical Laboratory(1957).

[4] F. Rosenblatt: *Principles of neurodynamics.* Spartan, New York (1962).

[5] M. Minsky and S. Papert: *Perceptrons.* MIT press, Cambridge, MA (1969).

[6] J. J. Hopfield: "Neural networks and physical systems with emergent collective computational abilities", *Proceedings of the National Academy of Sciences, USA* **79** (1982) 2554.

[7] T. Kohonen: *Self-Organization and Associative Memory.* Springer-Verlag, Berlin (1984).

[8] A. G. Barto, R. Sutton, and W. Anderson: "Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems", *IEEE Transactions on Systems, Man and Cybernetics* **13** (1983) 834.

[9] Rumelhart D. E., Hinton G. E., and Williams R. J.: "Learning representations by back-propagating errors", *Nature* **323** (1986) 533.

[10] URL: http://www.image-net.org/

[11] S. Hochreiter and J. Schmidhuber: "Long short-term memory", *Neural Computation* **9** (1997) 1735.

[12] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio: *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*, arXiv:1406.1078v3 (2014).

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner: "Gradient-based learning applied to document recognition", *Proc. of the IEEE* **86** (1998) 2278-2324.

[14] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio: *Generative Adversarial Nets*, Proceedings of Neural Information Processing Systems, (2014) 8.

[15] The report of the investigation for the Tesla S accident occurred on May 7, 2016 in Florida. U.S. Department of Transportation, National Highway Traffic Safety Administration (NHSA), URL: https://static.nhtsa.gov/odi/inv/2016/INCLA-PE16007-7876.PDF

[16] C. Szegedy et al.: *Intriguing properties of neural networks*, arXiv:1312.6199v4 (2014).

[17] Black Hat USA 2015 *The full story of how that Jeep was hacked.* `https://www.kaspersky.com/blog/blackhat-jeep-cherokee-hack-explained/9493/`

[18] I. J. Goodfellow, J. Shlens, and C. Szegedy: *Explaining and Harnessing Adversarial Examples*, arXiv:1412.6572v3 (2015).

Giorgio Buttazzo: è professore ordinario di Ingegneria Informatica presso la Scuola Superiore Sant'Anna di Pisa, dove insegna Real-Time Systems e Neural Networks. Attualmente si occupa di architetture e algoritmi per fornire un supporto predicibile agli algoritmi di controllo e all'Intelligenza Artificiale.

La lezione mancata

La meccanica statistica dei sistemi complessi

The more is different. The behavior of large and complex aggregates of elementary particle, it turns out, is not to be understood in terms of a simple extrapolation of a few particles. Instead, at each level of complexity, entirely new properties appear, and the understnading of the new behaviors requires research which I think is as fundamental in its nature as any other.

Philip Warren Anderson

Elena Agliari *Dipartimento di Matematica "Guido Castelnuovo", Sapienza Università di Roma*

Adriano Barra *Dipartimento di Matematica e Fisica "Ennio De Giorgi", Università del Salento*

In questo articolo esponiamo una prospettiva, stilizzata e sintetizzata, inerentemente la genesi della Teoria della Complessità nella Meccanica Statistica: dopo un'introduzione storica ed una minima digressione sulla Meccanica Statistica tout-court utilizziamo la stessa per riassumere alcune riflessioni sulla Teoria dei Sistemi Complessi, originariamente formulata da Giorgio Parisi, ed alcune sue implicazioni.

Come *overture*, per prendere confidenza amatoriale con la Meccanica Statistica, mostriamo che *complesso* e *riduzionistico* non sono necessariamente aggettivi antitetici: per fare questo tratteremo esempi di *complessità emergente* utilizzando unicamente modelli in cui le energie sono forme quadratiche nelle loro variabili microscopiche e, mediante il principio di massima entropia, legheremo queste funzioni costo al riduzionismo in un'accezione che chiameremo *riduzionismo statistico*.

Il cuore del lavoro che segue vuole invece enfatizzare come la conditio sine qua non per avere un comportamento complesso sia la simultanea presenza, nella pletora di interazioni tra gli elementi miscroscopici che formano il sistema in esame, tanto di coesione quanto di competizione (anche in un quadro riduzionista): a supporto di questa visione, e come omaggio alla prolifica e storica Scuola di Sistemi Dinamici del Fiorini, chiuderemo lo scritto mostrando alcune similitudini nei requisiti per la genesi della fenomenologia complessa in Meccanica Statistica (a dire un connubio tra coesione e competizione)

con gli ingredienti fondamentali per la genesi del chaos deterministico nei *sistemi dinamici*, traendo ispirazione, inerentemente questi ultimi, dalla mappa logistica di Robert May. Inizieremo questa chiacchierata informale usando variabili microscopiche continue, con cui si ha forse più familiarità nei primi anni di Università, per poi estendere il quadro a variabili discrete nella seconda metà dello scritto: l'idea sottostante è che, in questa maniera, il quadro tracciato possa avere un suo senso proprio ma possa anche essere usato, volendo, come una chiave di lettura per alcuni articoli del numero di Ithaca sull'Ingelligenza Artificiale.

Parte Uno: il Riduzionismo Statistico

Perchè nel corsi di base di Statistica o nei laboratori di Fisica ci è stato insegnato che, per descrivere opportunamente un campione statistico, abbiamo bisogno (almeno) di due quantificatori (a dire, la media e la deviazione standard)? E cosa c'entra questo con il telaio riduzionista che appare egemone nella Meccanica Elementare? Per rispondere a queste domande, serve una definizione minimale di Principio di Riduzione, sulla quale non c'è unanimità ma che, sovente, viene sintetizzata sancendo che, per conoscere la soluzione di un dato problema, possiamo spezzare lo stesso in sotto-problemi da risolvere separatamente ottenendo delle soluzioni parziali che, opportunamente sommate, forniscono la soluzione del sistema oggetto di studio iniziale. Questa operazione può essere schematizzata come *Se A implica C e B implica D, allora (A+B) implica (C+D)*

$$(A \to C) \wedge (B \to D) \Rightarrow (A \cup B) \to (C \cup D).$$

Un archetipo di comportamento riduzionista in Fisica, nella Meccanica Elementare, è offerto dall'*oscillatore armonico*: chiamata x una coordinata spaziale e k il valore della costante elastica della molla in questione (e.g., una bilancia) consideriamo una forza $F_k = -kx$ (scalare per semplicità). Se poniamo sulla bilancia una mela e la sua molla si estende per un tratto x_1, se ne ricava un valore per la forza peso $F_1 := F_k(x_1) = -kx_1$ (l'informazione sull'elongazione è convertita in un'informazione sul peso della mela). Parimenti, in un secondo esperimento, ponendo sulla stessa bilancia una pera, questa si estenderà per un tratto x_2, e se ne ricava una forza $F_2 := F_k(x_2) = -kx_2$; chiaramente, se pesiamo simultaneamente la mela e la pera, la bilancia si estenderà per un tratto $x_1 + x_2 = x_{totale}$ ottenendo una forza $-k(x_1 + x_2) = -kx_{totale} = F_1 + F_2 =: F_{totale}$. L'oscillatore armonico gioca un ruolo così importante in Fisica che quando, in un certo campo di studio, un modello costituisce il riferimento da cui discende una pletora di estensioni e variazioni sul tema, si epiteta lo stesso come *l'oscillatore armonico del campo in questione*.
Per i nostri fini è importante ora notare che, se la generica forza $F_k(x)$, nel generico parametro k, è una funzione lineare delle variabile microscopiche x, l'energia $E_k(x)$ a questa associata (ovvero il suo integrale lungo la traiettoria) è una forma quadratica nelle stesse variabili microscopiche, a dire, se $F_k(x) := -kx \Rightarrow E_k(x) = \frac{1}{2}kx^2$: quest'osservazione banale avrà il suo peso per capire cosa intendiamo per *riduzionismo statistico*.

Il Principio di Massima Entropia in Fisica ed in Matematica

Le prime formulazioni del *principio di massima entropia* sono state dipinte in Fisica da celebri mani quali quelle di Ludwig E. Boltzmann, Josiah W. Gibbs e James C. Maxwell alla volta di una formulazione meccanicistica e miscrosopica della Termodinamica sul finire della *belle époque* (in un periodo prolifico che ha portato anche alla nascita della Teoria Ergodica). Di contro, in Matematica (in particolare in Statistica), negli anni $'50$ del secolo scorso, Edwin T. Jaynes mostrava come un generico telaio inferenziale derivabile dal principio di massima entropia fosse in completa armonia con quelli derivabili mediante principi variazionali classici quali il *principio di massima verosimiglianza*.

Vediamo cosa dice il principio di massima entropia, partendo dalla sua formulazione come metodo matematico generale e successivamente come postulato cardine della Fisica. Per farlo introduciamo ed utilizziamo l'entropia nella sua espressione fornita da Claude Shannon nell'ambito della Teoria dell'Informazione (ma nella

prossima sottosezione vedremo come questa sia fondamentalmente la stessa definita dai fisici, rendendo Teoria dell'Informazione e Meccanica Statistica due volti dello stesso Giano).
Consideriamo N estrazioni indipendenti della variabile casuale $x \in \mathbb{R}$, estratte da una densità di probabilità incognita $p(x)$; per fissare le idee possiamo immaginare di avere a che fare con una moltitudine di N molle indipendenti[1] per cui $x_1, x_2, ..., x_N$ rappresentano le loro elongazioni in un opportuno sistema di riferimento. L'entropia di Shannon $S[p]$ associata a questo sistema si scrive

$$S_{\lambda_0}[p] = -\int_{-\infty}^{+\infty} p(x)\ln p(x)dx + \lambda_0\left(\int_{-\infty}^{\infty} p(x)dx - 1\right),$$

dove l'ultimo termine nel membro di destra dell'equazione esprime il vincolo (in forma di moltiplicatore Lagrangiano) che la variabile funzionale p che compare in $S_{\lambda_0}[p]$ sia proprio una probabilità (la relazione di chiusura).
Massimizzando questa entropia otteniamo come soluzione la distribuzione $p(x)$ meno strutturata possibile[2], quella cioè *al massimo disordine eventualmente in accordo con i dati a disposizione su questo insieme di molle*. In particolare, se non facciamo alcuna misura su queste molle, semplicemente ne postuliamo l'esistenza, estremizzando $S_{\lambda_0}[p]$ rispetto a p e λ_0 si ottiene una distribuzione uniforme. La derivazione è tanto facile quanto intuitivo il risultato: senza nessuna informazione sul sistema, il principio ci suggerisce che qualunque realizzazione ha la stessa probabilità di verificarsi; d'altronde, se non suppliamo al principio alcuna informazione sul sistema, cos'altro di ragionevole dovremmo aspettarci?
Supponiamo ora di aver effettuato alcune misure su queste molle, per esempio, assumiamo di conoscere il momento primo empirico (i.e., la media campionaria) $\langle x\rangle_{\rm exp}$ di $p(x)$ ed il momento secondo empirico $\langle x^2\rangle_{\rm exp}$ di $p(x)$ (e quindi la varianza empirica $\sigma^2_{\rm exp} := \langle x^2\rangle_{\rm exp} - \langle x\rangle^2_{\rm exp}$), ed imponiamo, sempre mediante vincoli Lagrangiani, che la distribuzione di massima entropia che vogliamo sia in accordo con queste misure, cioè tale che i primi due momenti, che chiamiamo teorici, riproducano quelli sperimentali, i.e. $\langle x\rangle_{\rm exp} = \langle x\rangle_{theor} := \int_{-\infty}^{+\infty} xp(x)dx$ e $\langle x^2\rangle_{\rm exp} = \langle x^2\rangle_{theor} := \int_{-\infty}^{+\infty} x^2p(x)dx$. Dobbiamo quindi estremizzare rispetto a p ed ai tre moltiplicatori Lagrangiani $\lambda := \{\lambda_0, \lambda_1, \lambda_2\}$ la seguente espressione

$$\begin{aligned} S_\lambda[p] &= S_{\lambda_0}[p] + \lambda_1(\langle x^1\rangle_{theor} - \langle x^1\rangle_{\rm exp}) + \\ &+ \lambda_2(\langle x^2\rangle_{theor} - \langle x^2\rangle_{\rm exp}). \end{aligned} \quad (1)$$

Risolvendo il sistema di quattro equazioni che tale estremizzazione implica otteniamo

$$\hat{p}(x) = \frac{1}{Z}e^{\frac{-(x-\hat{\lambda}_1)^2}{2\hat{\lambda}_2}}, \quad (2)$$

dove $\hat{\lambda}_1 = \langle x\rangle_{\rm exp}$, $\hat{\lambda}_1 = \sigma^2_{\rm exp}$ e la normalizzazione (nota in Meccanica Statistica come *funzione di partizione*) vale $Z = \sqrt{2\pi\sigma^2_{\rm exp}}$, cioè una distribuzione di Gauss, centrata proprio sul valor medio delle elongazioni misurate e con varianza che ne riproduce le fluttuazioni: un risultato che il Teorema del Limite Centrale avrebbe previsto senza necessità di fare alcun conto.

In Fisica, in particolare in Meccanica Statistica, il Principio di Massima Entropia si formula in maniera differente: considerato un generico sistema, descritto sempre mediante coordinate microscopiche x, ed al quale si assegna energia media $\bar{E}$, quello che vogliamo massimizzare si scrive

$$S_{\lambda_0,\lambda_E}[p] = S_{\lambda_0}[p] + \lambda_E\left(\int p(x)E(x)dx - \bar{E}\right).$$

Il moltiplicatore lagrangiano λ_E ha un ruolo cruciale: impone che l'energia media sia definibile (i.e., che esista), ovvero che il sistema possa termalizzare poiché così facendo rimuoviamo eventuali patologie, come divergenze iperboliche in cui particelle scappano via all'infinito (rimanendo perennemente in moto) o collassano reciprocamente (e.g., la legge di repulsione di Coulomb tra due protoni o la sua variante attrattiva tra un protone ed un elettrone).
La massimizzazione dell'entropia $S_{\lambda_0,\lambda_E}[p]$ ci restituisce la seguente espressione, nota come distribuzione di Maxwell-Boltzmann o di Gibbs:

$$p_{Gibbs}(x) = \frac{1}{Z}e^{-E(x)/T}, \quad (3)$$

dove Z ha sempre il ruolo di normalizzazione e T -la temperatura del *bagno termico* nel quale il sistema vive- è legata al moltiplicatore di Lagrange, che tacitamente mostra quanto sia esponenzialmente improbabile trovare il sistema in uno stato lontano da quello a cui corrisponde la sua minima energia.

Uniamo allora i punti elencati fino a qui: se co-

[1] Questo esempio si potrebbe rendere molto più interessante facendo interagire le molle, invece di considerare un "gas" di oscillatori, per esempio, senza ricorrere alla meccanica quantistica, ritrovando il calore specifico dei solidi empiricamente stimato essere costante (per temperature non troppo basse) da Dulong e Petit, ma esulerebbe dagli scopi del manoscritto.

[2] L'ottenere come risultato una distribuzione *il meno strutturata possibile* deve essere visto in una luce positiva, à la *rasoio di Occam*: sia meno *pregiudizi aprioristici* introduco meno rischio di sbagliare l'interpretazione dei risultati, sia meno *assumo in partenza*s più capisco alla fine (a parità di contenuto informativo nel risultato). Si vedano, a tal proposito, i contributi di Michele Castellana e Matteo Marsili, in questo volume.

struiamo una disciplina sul telaio riduzionista (e.g., la Fisica dell'oscillatore armonico) e quindi usiamo forze lineari generate da energie che sono forme quadratiche, approdiamo sempre a distribuzioni di massima entropia che sono Gaussiane nelle variabili microscopiche, per cui per estrarre informazioni dai relativi esperimenti sarà sufficiente controllare il momento primo ed il secondo (nel limite di cospicua statistica) come sancito dai Teoremi del Limite Centrale: è questa armonia di indagine, tra la Meccanica Elementare (per formulare modelli) e l'Inferenza Statistica di base (per verificarne la bontà) che, noi autori del presente scritto, chiamiamo *riduzionismo statistico*.
Chiaramente, cosí come esistono modelli meccanici che esulano da una trattazione riduzionista (e.g., i sistemi non lineari), esistono distribuzioni che si svincolano dai dictat dei teoremi di convergenza in probabilità (e.g., le distribuzioni a potenza): tanto i primi generano sovente caoticità nei Sistemi Dinamici quanto i secondi di prassi generano complessità nella Meccanica Statistica.

Il modello di Erhnenfest: l'entropia da Gibbs a Shannon

Poiché l'entropia è un'osservabile cardine dell'intera Meccanica Statistica, e poiché a breve vogliamo descrivere sistemi le cui variabili microscopiche sono interi e non reali [3], in questa sotto-sezione affrontiamo da una prospettiva entropica un modello discreto elementare, il modello di Erhenfest, che si preoccupa di *far termalizzare palline di due diversi colori in due urne (o buche)*.
Per vedere che legame esista tra le due celebri formule dell'entropia, quella supplita da Boltzmann $S = \ln \Omega(x)$ (dove $\Omega(x)$ è *il numero di microstati compatibili col macrostato di cui calcoliamo S*) e quella supplita da Shannon $S = -p(x) \ln p(x)$ (dove p è la probabilità di osservare un particolare microstato x), vediamo la termalizzazione del modello di Erhenfest, che, essendo un modello iso-energetico, ci permette proprio di prendere confidenza con l'entropia.
In questo modello esistono due buche, inizialmente contenenti la prima N_A palline rosse ed N_B palline blu, la seconda N_B palline rosse ed N_A palline blu, tali che $N = N_A + N_B$; le buche possono liberamente e casualmente scambiarsi palline tra di loro (le palline si mischiano quindi), conservando il numero N di palline in ciascuna buca. Ci chiediamo due cose:
Domanda di Statistica: quale è la distribuzione di massima entropia di questo sistema?
Domanda di Fisica: il sistema raggiunge l'equilibrio? Se sì, che caratteristiche ha?

Iniziamo dalla descrizione statica (quindi rispondendo alla domanda Statistica): il numero di configurazioni possibili per una buca di questo sistema (per l'altra il discorso è identico) è $\Omega = \binom{N}{N_A}$, pertanto l'entropia di Boltzmann di pertinenza può essere scritta come $S = \ln \Omega = \ln N! - \ln N_A! - \ln N_B! \sim N \ln N - N_A \ln N_A - N_B \ln N_B$. Possiamo ora aggiungere $0 = N_A + N_B - N$ al membro di destra della precedente equazione lasciandone ovviamente inalterato il contenuto informativo per ottenere $S = -N_A \ln(\frac{N_A}{N}) - N_B \ln(\frac{N_B}{N})$. Ora facciamo due operazioni, la prima banale la seconda lievemente meno: la prima introduce le densità di particelle $\rho_i := N_i/N, \;\; i \in (A, B)$, la seconda introduce anche la densità di entropia, cioè ci interessiamo ad S/N al posto di S. Cui prodest? Il vantaggio nel passaggio da quantità linearmente estensive (ricordiamo che le osservabili termodinamiche scalano linearmente con il volume, i.e. con il numero di particelle) a quantità intensive (basta dividere le estensive per il volume, qui rappresentato da N) risiede nel fatto che per le prime i valori medi crescono linearmente in N alla stregua delle loro varianze (non delle loro deviazioni standard!), mentre per le seconde, non potendo crescere in N per definizione, le relative distribuzioni si *concentrano* sul valor medio, in accordo con la convergenza a zero delle loro varianze campionarie per $N \to \infty$. Per queste osservabili, anche dette *automedianti*, si ottengono così delle comode rappresentazioni δ-like, molto usate nella Meccanica Statistica.
Notando che ρ_A e ρ_B, se interpretate come frequenze, per $N \to \infty$ tendono alle rispettive probabilità per la legge dei grandi numeri (infatti formano una partizione, i.e. $(\rho_A \cap \rho_B = 0) \wedge (\rho_A \cup \rho_B = 1)$), tralasciando le etichette A, B in favore di $\rho, 1 - \rho$, e chiamando $s[\rho] = S[\rho]/N$ l'entropia intensiva, possiamo scrivere

$$s[\rho] = -\rho_A \ln \rho_A - \rho_B \ln \rho_B = \rho \ln \rho + (1-\rho) \ln(1-\rho),$$

che molto somiglia alla rappresentazione à la Shannon per un esperimento bernoulliano $s[p] = p \ln p + (1-p) \ln(1-p)$, dove p rappresenta la "probabilità di successo".
Così, partendo dall'espressione di Boltzmann per l'entropia, siamo approdati a quella di Shannon.
È elementare verificare che il massimo di $S[\rho]$ (che si ottiene richiedendo $dS[\rho]/d\rho = 0$) si ha per $\rho = 1/2$, cioè quando un ugual numero di palline di ogni tipo riempie le buche, ovvero per la configurazione di *massimo disordine* (i due casi di massimo ordine sarebbero stati, di contro, tutte palline blu in una buca e rosse nell'altra e la situazione speculare con tutte le palline rosse nella prima e tutte quelle blu nella secon-

[3] Questo passaggio permette di usare questo scritto come prospettiva con cui leggere alcuni articoli pubblicati in parallelo sul volume dedicato all'Intelligenza Artificiale di Ithaca. Con la stessa filosofia, nel seguito, come prototipo di sistema semplice studieremo il modello di Curie-Weiss, che è un modello definito al discreto, ma si sarebbe parimenti potuto usare il modello di Van Der Waals, che è un modello definito nel continuo.

da). Questa configurazione (distribuzione informe) è anche quella prevista in uno scenario inferenziale nel caso in cui nessuna informazione empirica venga fornita al Principio di Massima Entropia.
Proseguiamo ora con una descrizione dinamica, rispondendo alla domanda di Fisica: che scenario offre la termalizzazione all'equilibrio?
Proviamo a scrivere un'evoluzione dinamica per la densità di palline, ma senza ricorrere a strumenti matematici sofisticati (e.g., le catene di Markov), scegliamo delle regole dinamiche che siano semplicemente *ragionevoli* (si veda anche figura 1):
1) Scegliamo una pallina a caso, uniformemente da una delle due buche, diciamo la prima per fissare le idee, quindi con probabilità di esser estratta pari ad $1/N$.
2) Attribuiamo a questa pallina una probabilità q di rimanere nella buca di appartenenza ed una probabilità $1-q$ di saltare nell'altra buca, secondo questo criterio:
$\Delta N_A = +1$ con probabilità $p = \frac{N_B}{N}(1-q)$ (N_A cresce quando una pallina blu si trasferisce nella seconda buca).
$\Delta N_A = -1$ con probabilità $p = \frac{N_A}{N}(1-q)$ (N_A decresce quando una pallina rossa lascia la prima buca).
Se ora scriviamo (oculatamente perché ci sono dei passaggi al limite che non svisceriamo come si dovrebbe) il rapporto incrementale $\Delta N_A/\Delta t$, dove il lasso di tempo da impiegarsi lo assumiamo ragionevolmente proporzionale al volume delle buche, cioè $\Delta t \sim N\delta t$, e parimenti assumiamo $\Delta N_A \sim N\delta\rho_A$, possiamo scrivere un'equazione differenziale ordinaria per ρ_A che si legge

$$\frac{\delta\rho_A}{\delta t} = \rho_B(1-q) - \rho_A(1-q) \to \frac{d\rho}{dt} = (1-q)(1-2\rho).$$

dove le stesse considerazioni fatte nella statica ci permettono di dimenticare le etichette A, B.
È immediato verificare che, a patto che la dinamica esista (i.e., $q \neq 1$), un rilassamento all'equilibrio avviene quando $d\rho/dt = 0$, cioè quando $\rho = 1/2$, in completo accordo con il previo approccio.

Parte Due (statica): Semplice versus Complesso in Meccanica Statistica

La distribuzione di probabilità che quindi usiamo per sancire l'occorrenza di una certa configurazione x di un dato sistema è $p(x) = e^{-\frac{E(x)}{T}}/Z$, dove $E(x)$ rappresenta l'energia del sistema nella configurazione x mentre T rappresenta la temperatura a cui il sistema vive: se la temperatura è molto alta (i.e., $T \gg E$), l'evoluzione del sistema è fondamentalmente randomica, infatti, è governata dalla massima entropia e

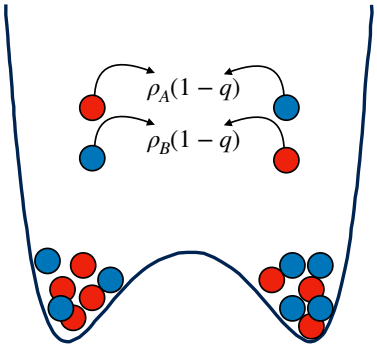


Figura 1: *Nel modello di Erhenfest si hanno due buche (o urne), la prima contenente N_A palline rosse ed N_B palline blu, la seconda contenente N_B palline rosse ed N_A palline blu; nell'esempio mostrato in figura la prima buca è quella di sinistra ed $N_A = 4$, $N_B = 3$. Il sistema viene fatto "termalizzare" iso-energeticamente, conservando cioè il numero $N = N_A + N_B$ di palline in ciascuna buca. La probabilità che una pallina rossa si trasferisca dalla prima alla seconda buca (e quindi, simultaneamente, una pallina blu si trasferisca dalla seconda alla prima buca) è $(N_A/N)(1-q)$ e, come conseguenza di questo scambio, il numero di palline blu nella prima urna cresce di una unità. Lo scambio opposto, per cui il numero di palline rosse nella prima urna cresce di una unit; a si verifica con una probabilità $(N_B/N)(1-q)$.*

le varie configurazioni tendono ad essere equibrobabili, come si evince banalmente prendendo il limite $T \to \infty$ di $p(E)$[4]; di contro, nel limite $T \to 0$, il contributo entropico viene soppresso e la dinamica del sistema torna ad essere una deterministica ricerca del minimo dell'energia, alla volta dell'ordine celato dietro la formula che esplicita la forma funzionale di E (che funge da funzione di Lyapounov a temperatura nulla grazie al Teorema Spettrale) in termini delle configurazioni microscopiche x: capire questo barcamenarsi tra ordine e disordine è il ruolo della Meccanica Statistica.
Una questione cardine nella Meccanica Statistica di una cinquantina di anni fa verteva sullo studio delle transizioni di fase *classiche*: se a temperature alte vince il disordine e la massima entropia è dominante nella selezione delle configurazioni osservabili, mentre alle basse temperature vince l'ordine e a dominare è la minima energia, due osservazioni sono facilmen-

[4] C'è una ben nota relazione termodinamica tra l'aumento della temperatura e quello dell'entropia e, proprio come l'entropia ha un ruolo tanto nella Fisica quanto nella Teoria dell'Informazione, lo stesso succede per la temperatura: si veda a tal proposito il box *la temperatura ubriaca*.

te deducibili:
a) per capire "chi prevalga su chi" è bene confrontarle, energia ed entropia, ed effettivamente i meccanici statistici studiano le proprietà dell'energia libera F di un sistema, definita come $F := E - TS$
b) tra il limite di temperatura infinita, dove sicuramente vince la massimizzazione di entropia, e quello di temperatura nulla, dove sicuramente vince la minimizzazione dell'energia, qualcosa deve succedere: quel qualcosa è una *transizione di fase*.
Di cosa si tratta? All'atto pratico è evidente che la conoscenza della fase di un sistema è importante per una moltitudine di motivi. Si pensi ad una bottiglietta d'acqua che vogliamo bere: sapere che l'acqua sia nella sua fase liquida piuttosto che solida certamente ci evita una figuraccia, qualora cercassimo di far uscire un cilindro di ghiaccio dall'esile buco della bottiglietta in presenza di amici; l'altro limite sarebbe ancor peggiore: provare ad aprire una bottiglietta di vapore per tentare di bere non produrrebbe neanche più una risposta ilare da amici ormai attoniti...
Il nodo centrale è che, qualsiasi sia il nostro grado di conoscenza inerentemente la molecola dell'acqua (conosciamo la chimica e sappiamo che ci sono due atomi di idrogeno per atomo di ossigeno; di più, conosciamo la meccanica quantistica ed addirittura possiamo dire che gli orbitali formano degli angoli medi di 104.4 gradi tra loro, etc.), da tutta questa informazione (che neanche serve) in nessuna maniera si evince che quando un bicchiere d'acqua si mette in freezer questo ghiaccia e parimenti quando si mette una pentola d'acqua sul fuoco questa evapora formando delle caratteristiche bolle di vapore all'interno della rimanente acqua in forma liquida. Se abbiamo perseguito lo schema logico di riduzione spezzando il sistema *bicchiere d'acqua* in sotto-sistemi *le singole molecole* e le abbiamo studiate *indipendentemente* (trascurando cioè il loro interagire), la nostra ignoranza sull'emergenza di una transizione di fase è alquanto ragionevole: le transizioni di fase sono proprietà collettive, *fenomeni emergenti dalle reti che le molecole d'acqua formano mediante il loro tumultuoso interagire* nelle quali, tipicamente, le distribuzioni statistiche non rispettano i dictat del Teorema del Limite Centrale, optando per comportamenti scale-free[5]. Notiamo tuttavia che questo tipo di transizioni di fase, benché estremamente interessanti, non rappresentano dei veri "fenomeni complessi", nel senso che le fasi che separano possono essere estrapolate direttamente a partire dalla forma funzionale dell'energia $E(x)$ e corrispondono a diversi gradi di ordine (o simmetria) facilmente definibile. Per chiarire meglio questi concetti, nella prossima sottosezione approfondiamo un esempio di sistema *semplice*, prima di avventurarci alla volta di un sistema *complesso* nella successiva sottosezione.
Concludiamo sottolineando che, in questo scritto, volutamente continueremo a discutere solo energie che siano forme quadratiche nelle loro variabili microscopiche (e quindi, in linea di principio, riducibili) al fine di mostrare che non è così banale sancire un divario tra lo schema di riduzione e la complessità.

[5]Un esempio di distribuzione a potenza, tipico delle transizioni di fase, lo si ottiene guardando l'istogramma della grandezza delle bolle di vapore che risalgono l'acqua *bollente*: tipicamente queste spannano su molti ordini di grandezza, dal microscopico all'agevolmente visibile ad occhio nudo.

L'oscillatore armonico dei sistemi semplici: il Curie-Weiss

Assunto ormai che, con il termine alquanto lasco *sistema* intendiamo una generica collezione di oggetti, i quali possono eventualmente interagire, in questo testo optiamo per prendere "a paradigma del semplice" il *modello di Curie-Weiss*: questo modello, inizialmente proposto da Lenz, costituisce l'oscillatore armonico delle transizioni di fase paramagnete-ferromagnete nei sistemi composti da spin (è quindi un modello definito sugli interi) e le due fasi sono chiamate *paramagnetica* (o *ergodica*, quella ad alta temperatura) e *ferromagnetica* (o *non-ergodica*, quella a bassa temperatura).
Vediamo di cosa si tratta: innanzitutto assumiamo che ci siano N spin il cui stato è indicato con $\sigma_i = \pm 1$, $i \in (1, ..., N)$ che interagiscono tutti con tutti e con la stessa intensità (i Fisici di solito chiamano questa approssimazione *di campo medio*, secondo una lunga tradizione iniziata da Llewellyn Thomas ed Enrico Fermi, mentre i Matematici chiamano le funzioni energia di questa classe *funzioni somma* di Aleksandr Khinchin). In virtú di questa interazione, quantificata mediante una costante di accoppiamento $J_{ij} > 0$ $(i, j = 1, ..., N)$ nell'eq. (4), gli spin cercano di allinearsi tra loro: maggiore J_{ij} e maggiore la propensione degli spin i e j al reciproco allineamento. Gli spin possono anche sentire il mondo esterno che viene loro supplito in forma di un campo magnetico $h \in \mathbb{R}$, che, se positivo, tenderà a far allineare gli spin verso lo stato $+1$, se negativo, verso lo stato -1. È allora ragionevole scrivere l'energia del modello di Curie-Weiss

La temperatura ubriaca

Abbiamo visto che, per un generico sistema di energia E, la sua distribuzione di equilibrio si legge $P(E) = e^{-E/T}/Z$, ma fino ad ora non ci siamo soffermati sul ruolo chiave della temperatura: sappiamo che un incremento di entropia si manifesta sovente trasportato dal calore (non per caso quando ascoltiamo musica a volume sostenuto e mettiamo la mano sull'amplificatore che la produce, questo si è scaldato) ma il legame tra temperatura e disordine è ancora oscuro. È immediato sincerarsi che se $T >> E$ fondamentalmente la distribuzione di Boltzmann pesa uniformemente qualunque stato accessibile al sistema (mentre nel limite $T \to 0$ il sistema si impunta sull'unico stato di minima energia), quindi effettivamente T deve giocare un ruolo cruciale, e molto generale.
Come fa la temperatura, un'osservabile in linea di principio relegabile alle scienze applicate (fisica, chimica, etc.), ad avere un ruolo così saliente anche in un generale telaio inferenziale *à la* Yajnes (che si usa, per dire, anche in Economia ed in Intelligenza Artificiale)? Stiamo chiamando *temperatura* un concetto piu' generale? Si.
La temperatura obbedisce l'equazione di Fourier, un'equazione alle derivate parziali dove il termine di derivata temporale è del primo ordine e quindi, se si applica il *time reversal*, si manda cioè $t \to -t$, il quadro che lei dipinge cambia completamente, cioè T percepisce la freccia del tempo (cosa che non avviene per esempio nella propagazione delle onde, fruendo i Dalambertiani anche di derivate temporali seconde e non prime).
L'equazione di Fourier si può ottenere mediante un attento passaggio al limite di un modello discreto chiamato *random walk* (o *drunk walk* per i più epicurei) che si può riassumere come segue: immaginiamo che il camminatore viva su una retta (discretizzata) e, ad ogni scoccare di orologio (ad esempio puntualmente ogni secondo), questi si muove a passi obbligati, con probabilità $1/2$ verso destra e $1/2$ verso sinistra. Nel *passaggio al continuo*, possiamo identificare la densità di probabilità $p(x,t)$ di trovare il camminatore nel punto x al tempo t proprio come la temperatura $T(x,t)$.
Se si prende questo camminatore come ragionevole modello di generatore di casualità, effettivamente capiamo che ad alte temperature lo scenario è dominato dal massimo disordine ed il caso favorisce la massima entropia, di contro diminuendo la temperatura, il principio di minima energia torna ad essere egemone ed il rigore scolpito nell'espressione dell'energia che di volta in volta si studia si manifesta mascroscopicamente ordinando il sistema.

così[6]

$$\begin{aligned} E(\sigma|J,h) &= -\frac{1}{N}\sum_{i<j}^{N,N} J_{ij}\sigma_i\sigma_j - h\sum_{i=1}^{N}\sigma_i \\ &\sim -\frac{NJ}{2}m^2 - Nhm, \end{aligned} \tag{4}$$

dove $m := N^{-1}\sum_{i=1}^{N}\sigma_i$ (con fortuito opportunismo, m può rappresentare tanto la *media atitmetica*, i.e., uno stimatore ottimale per l'inferenza statistica, per i Matematici quanto la *magnetizzazione* per i Fisici) rappresenta il *parametro d'ordine* del modello, a dire, un'osservabile in grado di discernere in quale fase viva il sistema (vide infra). Si noti che il fattore N che compare al denominatore del termine che esprime l'accoppiamento tra gli spin serve a garantire la lineare estensività della termodinamica, infatti, l'energia $E(\sigma|J,h)$ risulta scalare come un fattore N^1 che moltiplica oggetti che non crescono in N (seconda riga dell'eq. 4).
Chiaramente m è un parametro d'ordine poichè, se il sistema si trova nel regime ergodico (dominato dal caso imposto dalla massima entropia), per il Teorema del Limite Centrale (per un sistema sufficientemente

[6] Questa ipotesi è chiamata di campo medio ed è alquanto irragionevole in Fisica, dove le interazioni scemano con la distanza, ma parimenti di largo uso in altre branche della Scienza, quali, per esempio, l'Intelligenza Artificiale [1] e la Sociologia Quantitativa [2]. D'altronde è matematicamente conveniente rispetto ad un modello più realistico in Fisica, come per esempio un modello reticolare $E = \sum_{i=1}^{N} J_i\sigma_i\sigma_{i+1}$ in cui gli accoppiamenti coinvolgono solo spin contigui. Infatti, mentre nel campo medio le somme su $i \in 1, ..., N$ permettono di applicare Teoremi di Convergenza nel limite di volumi grandi, queste comodità si perdono nei reticoli, dove, anche se $N \to \infty$, i primi vicini (metrici) rimangono sempre gli stessi.

grande) risulta $m = 0$[7], di contro, se la temperatura permette agli spin di sentirsi reciprocamente (e quindi alla minima energia di imporsi), questi si allineano, risultando quindi in $m \neq 0$ (il lettore che non ha confidenza con l'energia sopra scritta spenda dieci secondi per sincerarsi che il termine di accoppiamento a due corpi $\sum_{ij} J_{ij}\sigma_i\sigma_j$ tende a far allineare gli spin grazie al segno meno davanti ad esso, assunto $J_{ij} > 0$).
La fase ad alta temperatura è relativamente semplice da caratterizzare anche con l'opportuno rigore formale, mentre cesellare la fase a bassa temperatura di questo sistema è un poco più complicato. Qui, senza perderci in tecnicismi, capiamo intuitivamente cosa succede per $T \to 0$: gli spin tendono ad allinearsi ed il panorama energetico (*landscape*) è dato da due buche di potenziale, una con il minimo in $+m$ e l'altra con il minimo in $-m$ le quali, al crescere della taglia del sistema N, non proliferano ma semplicemente diventano via via più profonde, in maniera da rompere rigorosamente l'ergodicità intrappolando il sistema in una sola delle due buche, si veda figura 2, nel limite termodinamico $N \to \infty$.
Siamo arrivati a proporre una nostra prima definizione di sistema semplice: *Un sistema termodinamico è semplice quando il numero di minimi che compongono il suo panorama energetico (intendendo con "energia" l'energia libera a sua disposizione ovviamente) non cresce come funzione del suo volume N (i.e. del numero di particelle che lo compongono)*, proprio come nel Curie-Weiss (o nel VanDerWaals). Come stiamo per vedere, questo non vale nei sistemi complessi, cosa che avrà conseguenze piuttosto significative sulla loro Fisica.

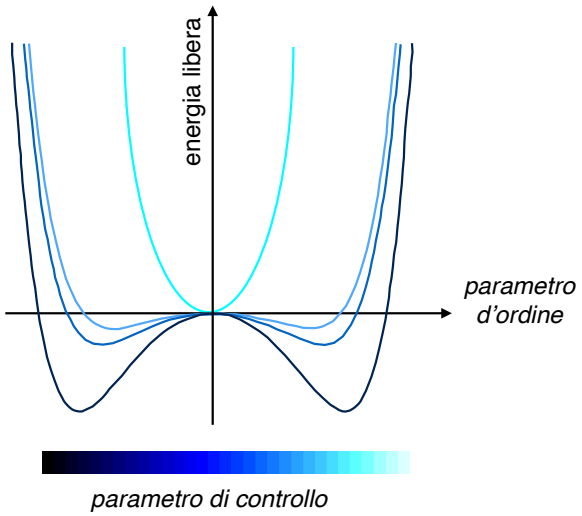


Figura 2: *Rappresentazione schematica dell'energia libera in funzione del parametro d'ordine, al variare del parametro di controllo per un sistema semplice. Per il modello di Curie-Weiss descritto nel testo, m gioca il ruolo di parametro d'ordine e la temperatura T gioca il ruolo di parametro di controllo. Si noti che a basse temperature il profilo energetico presenta due buche, mentre ad alte temperature si ha un unico minimo corrispondente alla fase ergodica.*

L'oscillatore armonico dei sistemi complessi: lo Sherrington-Kirkpatrick

A quasi un secolo di distanza, avendo investito incredibili sforzi nell'ottenere una Teoria Ergodica (sulla quale sorvoliamo totalmente, ma è una branca stupenda tra la Fisica e la Matematica), la rottura di ergodicità ci ha lasciati affascinati e sorpresi anche nei casi elementari quali il Curie-Weiss. D'altra parte, la nostra generazione, mentre studiava sui libri la rottura di simmetria per inversione di spin, ha vissuto come contemporanea la rottura di simmetria di replica, i.e. la rottura di invarianza permutazionale (che affrontiamo ora come paradigma dei sistemi complessi) e rimane sbalordita dalla stessa. In effetti, nonostante il fascino innegabile delle transizioni di fase, in realtà le proprietà emergenti del Curie-Weiss non sono poi così impreviste, poiché di fatto sono derivabili direttamente dalla forma dell'energia, si tratta solo di mettere gli spin in condizione di riuscire a sentirsi, cioè di abbassare il rumore *temperatura* opportunamente; al contrario, nei sistemi complessi il comportamento emergente non è immediatamente deducibile guardando alle interazioni microscopiche delle reti di elementi che li compongono.
A seguire discutiamo la meccanica statistica dei sistemi complessi -in gergo tecnico chiamati *vetri di spin* (spin glasses)- rubando maldestramente dalla Teoria di Giorgio Parisi [3] (e dalle successiva formulazione di Francesco Guerra [4] e Michel Talagrand [5]), e nell'ultima sezione dedicata ai Sistemi Dinamici, mostriamo come lo stesso passaggio logico (dal considerare reti di elementi solo coesivi al considerare reti in cui i costituenti elementali siano sia in coesione che in competizione) ci porti dagli equilibri di Malthus al caos deterministico [6].

Il primo concetto, che è anche quello cardine, che dobbiamo affrontare è quello della *frustrazione* (sostantivo preso in prestito, a ragione, dalla Psicologia, poiché nelle reti oggetto della presente digressione gli spin ricevono istruzioni conflittuali, rendendoli *frustrati*, si veda figura 3): nel Curie-Weiss, c'è solo *coesione* per cui gli spin lavorano sinergicamente per un bene comune, raggiungere la minima energia, l'unica minima energia possibile; cosa cambia quando questi devono anche competere, a dire, quando

[7] Ad alta temperatura l'energia non gioca un ruolo saliente quindi dovremmo aspettarci di riscontrare similitudini con il modello di Erhenfest: infatti, in un'analogia $1:1$ tra palline blu e spin up e palline rosse e spin down, nel previo paragrafo trovavamo un ugual numero di palline in ogni buca, a dire $m = 0$.

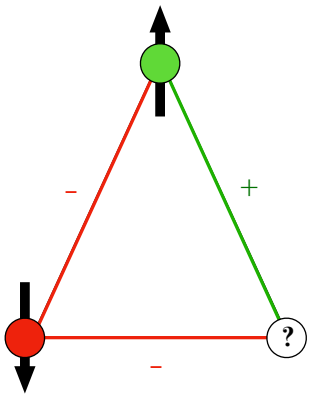


Figura 3: *Questa figura, beché essenziale, permette di capire l'origine della frustrazione: gli archi rossi indicano un accoppiamento negativo, quello verde un accoppiamento positivo. Se lo spin in alto si allinea verso l'alto, lo spin in basso a sinistra sollecitato ad orientarsi verso il basso ma così per allo spin in basso a destra arriveranno due stimoli contrastanti, rendendolo frustrato.*

oltre ad avere accoppiamenti positivi $J > 0$ (ferromagnetici) che favoriscono l'allineamento, si hanno anche accoppiamenti negativi $J < 0$ (antiferromagnetici) che favoriscono l'antiallineamento e quindi generano competizione? Per sincerarci di lavorare nel regime più tumultuoso possibile, prenderemo gli accoppiamenti tra spin casuali, identicamente ed indipendentemente estratti da una Gaussiana $\mathcal{N}[0,1]$, in maniera tale che si abbiano grossomodo la stessa quantità di *clausole*[8] allineanti (e.g., $J_{14} = +1$, che vuole mettere paralleli σ_1 e σ_4) e di *clausole* anti-allineanti (e.g., $J_{23} = -1$, che vuole mettere anti-paralleli σ_2 e σ_3): non con poco stupore, vedremo che da questo scenario massimamente disordinato in realtà, come fenomeno emergente e spontaneo scaturisce un ordinamento supremo, l'*ultrametricità*.

Un'occhiata alla figura 4 ci fa capire subito come in questo caso il numero di minimi (tecnicamente stati stabili e metastabili) proliferi con il volume (addirittura crescendo *esponenzialmente in* N[9]): un modello

[8] Quelli che in Fisica si chiamano *accoppiamenti* J_{ij} sono anche detti *archi* nella teoria dei grafi, o *clausole* in problemi di complessità algoritmica, o *sinapsi* (i.e. efficacie sinaptiche) nelle reti neurali o *weights* in machine learning e molto altro ancora, in ragione del campo di applicazione della meccanica statistica come telaio inferenziale di Jaynes.

[9] Se consideriamo spin binari, come quelli di Ising che stiamo usando, per un sistema di N spin esistono 2^N possibili configurazioni: un numero astronomico anche per N relativamente piccolo. Di contro è rimarchevole che, mettendoci a $T = 0$ per semplicità, il numero di minimi del sistema -se complesso- cresca parimenti esponenzialmente. Nei sistemi complessi, una fetta non così esigua dello spazio delle configurazioni è una soluzione, un compromesso di meta-stabilità... uno scenario alquanto diverso dal Curie-Weiss.

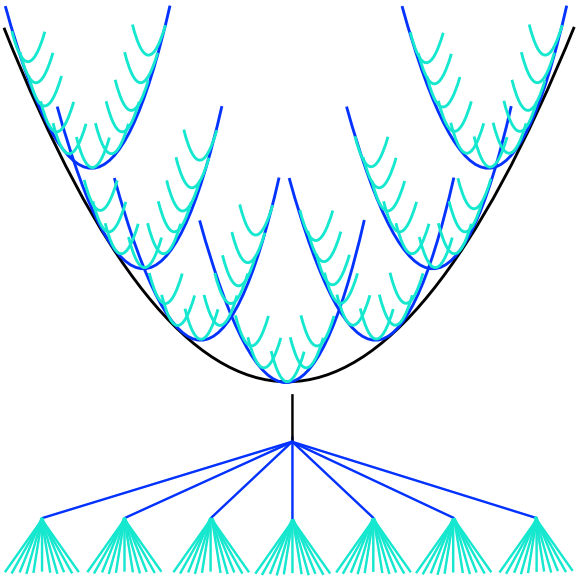


Figura 4: *Rappresentazione schematica dell'energia libera di un sistema complesso (in alto) e della struttura ultrametrica (in basso). Nel grafico in alto si osserva come ogni buca dia luogo (al livello di rottura di simmetria di replica successivo) ad un proliferare di altre buche in essa annidate (quasi come se gli steps di RSB fossero ingrandimenti di un microscopio) e questo corrisponde, nel grafico in basso, ad avere una foliazione (non nell'accezione geologica del termine) di repliche (le quali numerano le pendici degli ultimi rami disegnati in basso) che sono* identiche a gruppi*: repliche nello stesso gruppo (foglie appartenenti allo stesso ramo) sono identiche ma, man mano, che si allontanano reciprocamente nell'albero, diventano via via diverse.*

siffatto, sempre di campo medio, è stato introdotto da Sherrington e Kirkpatrick nel 1975 prendendo da loro il nome[10], e rappresenta l'oscillatore armonico dei Sistemi Complessi nella Meccanica Statistica ed ha energia

$$E_{SK}(\sigma|J) = -\frac{1}{\sqrt{N}} \sum_{i<j}^{N,N} J_{ij}\sigma_i\sigma_j \sim \frac{N}{2}(1 - q_{ab}^2)$$

con $P(J_{ij})$ (i.e. la probabilità di estrarre un valore dell'interazione tra gli spin i e j) assunta Gaussiana ed indipendente ed identicamente distribuita per tutti gli accoppiamenti. Il parametro d'ordine q_{ab} che compare nell'ultima espressione al membro di destra,

[10] Così come il modello di Curie-Weiss costituisce la versione campo-medio del modello di Ising-Lenz, il modello di Sherrington-Kirkpatrick costituisce la versione campo-medio del modello di Edwards-Anderson

chiamato *overlap* (o parametro d'ordine di Edward-Anderson) è meno intuitivo della magnetizzazione. Proviamo a capirci qualcosa: innanzitutto che si sia alle alte o alle basse temperature, dal punto di vista del previo parametro d'ordine $m := N^{-1} \sum_{i=1}^{N} \sigma_i$ poco cambia, questo è zero in entrambe le fasi. Quello che cambia è che, mentre nella fase ergodica (ad alta temperatura), gli spin si muovono continuamente e a casaccio, nella fase vetrosa (a bassa temperatura) gli spin rimangono congelati, ma sempre disordinatamente, quindi un'osservabile intelligente potrebbe essere la seguente: replichiamo il sistema, cioè, chiamata la copia originale del sistema *replica a* costruiamo una *replica b*, a dire una rete di spin che presenti la stessa identica realizzazione degli accoppiamenti della prima replica e vediamo quanto le rispettive configurazioni di spin, costrette ad obbedire alle stesse clausole, si assomigliano al variare del rumore (cioè della temperatura). Chiaramente la prospettiva di una soluzione unica è un miraggio: c'è un proliferare di compromessi più o meno buoni con cui il sistema dovrà barcamenarsi (come nella vita di tutti i giorni delle reti sociali), ma confrontare due configurazioni dovrebbe poter essere facile, in linea di principio basta fare il prodotto scalare tra queste e leggerne il valore, che null'altro è che l'overlap definito come $q_{ab} := N^{-1} \sum_{i=1}^{N} \sigma_i^a \sigma_i^b$. Questo sarà nullo a temperatura alta (poiché gli spin oscillano in continuazione in maniera scorrelata) mentre sarà diverso da zero a bassa temperatura, suggellando una certa similitudine tra i compromessi intrapresi dalla prima replica, rispetto a quelli scelti dalla seconda. . Come l'interazione tra spin nel Curie-Weiss rompe la simmetria di spin-flip a bassa temperatura, l'interazione tra repliche nello Sherrington-Kirkpatrick (che si vede con conti non elementari che qui non mostriamo), una volta raggiunta la temperatura critica, rompe la simmetria di replica: qui si ha la genesi della complessità, un'orgoglio in gran parte italiano (tralasciando numerosi contributi importanti, in estrema sintesi, la fenomenologia che stiamo per stilizzare è stata infatti prima euristicamente -ma esaustivamente- proposta da Giorgio Parisi in una serie di lavori conclusasi nel 1980 [3] e poi de facto dimostrata rigorosamente da Francesco Guerra in una serie di lavori conclusasi nel 2003 [4]).

La prima osservazione da fare, quando si rompe la simmetria di replica, è che il parametro d'ordine non è più automediante. Per capire questo concetto confrontiamo le distribuzioni del parametro d'ordine nel Curie-Weiss e nello Sherrington-Kirkpatrick nel limite termodinamico; indicando con $\bar{m}$ e con $\bar{q}_{ab}$ i rispettivi valori d'aspettazione, per il primo si ha

$$\lim_{N\to\infty} P_N(m) = \begin{cases} 0 & \text{se } m = 0 \\ 1 & \text{se } m = \pm\bar{m} \end{cases} , \qquad (5)$$

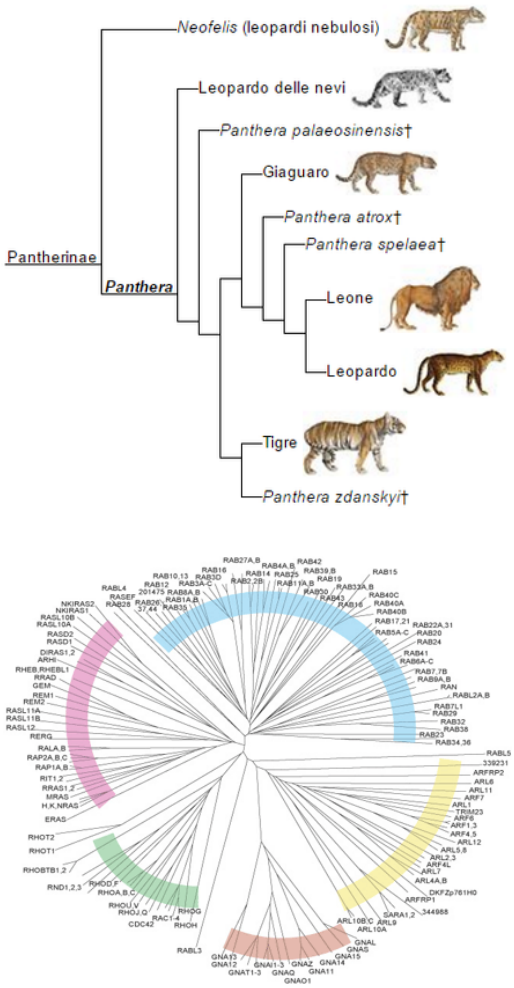


Figura 5: *Esempi di sistemi ultrametrici. Dall'alto: cladogramma basato su uno studio filogenetico di V.A. Vero, et al. (2018) su felini viventi e fossili e che indica la relazione delle specie Pantherinae (una sottofamiglia di felidi); Dendrogramma della superfamiglia delle proteine RAS ottenuto da J. Colicelli (2011) (le lunghezze dei bracci sono direttamente proporzionali al numero delle differenze tra le sequenze confrontate).*

perché più il sistema diventa grande, più i due minimi centrati su $\pm|\bar{m}|$ diventano profondi, mentre per il secondo

$$\lim_{N\to\infty} P_N(q_{ab}) \neq \begin{cases} 0 & \text{se } q_{ab} = 0 \\ 1 & \text{se } q_{ab} = \pm\bar{q}_{ab} \end{cases} , \qquad (6)$$

perché nei vetri di spin, più N cresce più le possibili soluzioni prolificano.

La seconda osservazione (legata ovviamente alla perdita dell'automedia per q_{ab}) riguarda il tipo di rottura e la scoperta dell'ultrametricità come proprietà di ordinamento spontaneo emergente[11]: le repliche non

[11] Uno spazio è *ultrametrico* se in esso la disuguaglianza triangolare è rafforzata in $d(x,z) \quad \leq$

sono tutte uguali o, in altre parole, la soluzione che prevede che tutte assumano gli stessi compromessi, la soluzione *replica simmetrica*, è sbagliata e, per esempio, la distribuzione congiunta di tre repliche si scrive

$$P(q_{12}, q_{13}) = \frac{1}{2}P(q_{12})\delta(q_{12} - q_{13}) + \frac{1}{2}P(q_{12})P(q_{13}),$$

cioè con probabilità $1/2$ gli overlap tra due repliche si comportano in maniera analoga (i.e. $q_{12} = q_{13}$, le repliche appartengono quindi allo stesso gruppo) e con probabilità $1/2$ si comportano in maniera indipendente (la loro probabilità fattorizza e le repliche appartengono a gruppi diversi, si veda la foliazione riportata in Figura 4).
Inoltre, quando si rompe una simmetria, se esiste un sotto-gruppo proprio della stessa, il sistema si decompone rispettandolo: quando passiamo dalla fase di alta temperatura dello spin glass (che è replica simmetrica) alla fase vetrosa, nel rompersi la simmetria di replica (l'invarianza di permutazione) si approda ad un sottogruppo che è ancora invariante sotto permutazione: questa è l'unica simmetria che si può rompere (gerarchicamente) infinite volte (nel limite termodinamico), come rappresentato dalle matrici a blocchi di Parisi a seguire (rispettivamente replica simmetrica Q_{RS}, ad uno step di rottura di simmetria di replica Q_{1RSB} e a due steps di rottura di simmetria di replica Q_{2RSB}). Con un volo pindarico, in maniera simile avviene il continuo biforcarsi delle possibilità di valori permessi dalla soluzione della mappa logistica di May (che tratteremo nella prossima sezione) al variare del suo parametro di controllo[12].

$$Q_{\text{RS}} = \begin{bmatrix} 1 & q_1 & q_1 & q_1 & q_1 & q_1 & q_1 & q_1 \\ q_1 & 1 & q_1 & q_1 & q_1 & q_1 & q_1 & q_1 \\ q_1 & q_1 & 1 & q_1 & q_1 & q_1 & q_1 & q_1 \\ q_1 & q_1 & q_1 & 1 & q_1 & q_1 & q_1 & q_1 \\ q_1 & q_1 & q_1 & q_1 & 1 & q_1 & q_1 & q_1 \\ q_1 & q_1 & q_1 & q_1 & q_1 & 1 & q_1 & q_1 \\ q_1 & q_1 & q_1 & q_1 & q_1 & q_1 & 1 & q_1 \\ q_1 & q_1 & q_1 & q_1 & q_1 & q_1 & q_1 & 1 \end{bmatrix},$$

$$Q_{\text{1RSB}} = \left[\begin{array}{cccc|cccc} 1 & q_2 & q_2 & q_2 & q_1 & q_1 & q_1 & q_1 \\ q_2 & 1 & q_2 & q_2 & q_1 & q_1 & q_1 & q_1 \\ q_2 & q_2 & 1 & q_2 & q_1 & q_1 & q_1 & q_1 \\ q_2 & q_2 & q_2 & 1 & q_1 & q_1 & q_1 & q_1 \\ \hline q_1 & q_1 & q_1 & q_1 & 1 & q_2 & q_2 & q_2 \\ q_1 & q_1 & q_1 & q_1 & q_2 & 1 & q_2 & q_2 \\ q_1 & q_1 & q_1 & q_1 & q_2 & q_2 & 1 & q_2 \\ q_1 & q_1 & q_1 & q_1 & q_2 & q_2 & q_2 & 1 \end{array}\right],$$

$$Q_{\text{2RSB}} = \left[\begin{array}{cc|cc|cc|cc} 1 & q_3 & q_2 & q_2 & q_1 & q_1 & q_1 & q_1 \\ q_3 & 1 & q_2 & q_2 & q_1 & q_1 & q_1 & q_1 \\ \hline q_2 & q_2 & 1 & q_3 & q_1 & q_1 & q_1 & q_1 \\ q_2 & q_2 & q_3 & 1 & q_1 & q_1 & q_1 & q_1 \\ \hline q_1 & q_1 & q_1 & q_1 & 1 & q_3 & q_2 & q_2 \\ q_1 & q_1 & q_1 & q_1 & q_3 & 1 & q_2 & q_2 \\ \hline q_1 & q_1 & q_1 & q_1 & q_2 & q_2 & 1 & q_3 \\ q_1 & q_1 & q_1 & q_1 & q_2 & q_2 & q_3 & 1 \end{array}\right].$$

In relazione alla poliedricità delle applicazioni della meccanica statistica complessa, le conseguenze interpretative della rottura di simmetria di replica sono molto profonde: un sistema (i.e., une rete di elementi potenzialmente in interazione) senza la minima organizzazione a-prioristica, anzi lasciato a termalizzare nalla sua anarchia (ma nel quale sia permesso ai suoi sudditi tanto di essere sinergici tra loro quanto, alla bisogna, di competere), si ordina spontaneamente secondo un ordinamento tassonomico (come l'ordinamento sociale sovrano, vassallo, valvassino, valvassore, etc, oppure, secondo l'ordinamento biologico degli alberi filogenetici, o semplicemente rispettando la catalogazione aristotelica della realtà in genere-specie[13], si veda figura 5).
Come accennato in precedenza, questi oggetti si organizzano in valli gerarchicamente annidate l'una nell'altra, come mostrato in Figura 4, e la densità di queste buche è esponenziale, cioè esiste una moltitudine di piccoli e fragilissimi minimi, poi molti minimi relativamente più profondi, poi sempre meno minimi ancora più stabili e così via. Se stimiamo con $e^{E_k/T}$ il tempo di uscita da una buca di energia E_k a causa di una fluttuazione termica (il *tempo di Arrenhius*), effettivamente, in questa perenne termalizzazione, diventa sempre più difficile procedere: man mano che si visitano buche poco profonde (che sono le più abbondanti e dalle quali è relativamente facile uscire), ci si annida in buche più profonde. Da un lato questo è bene, poichè rispetto alle buche meno profonde ormai alle spalle queste sono più stabili, dall'altro è un male perchè ambire a buche ancora più profonde tende ad essere via via illusorio.
Con una densità degli stati esponenziale, un qualunque calcolo naïve di meccanica statistica smette di convergere quando $T \to T_c^+ \equiv 1$:

$$\bar{E} := \int E\rho(E)e^{-\frac{E}{T}}dE = \int Ee^{E}e^{-\frac{E}{T}}dE,$$

quando $T = 1$ la densità degli stati si elide con l'esponenziale di Maxwell-Boltzmann, pertanto, il sistema rimane perennemente fuori dall'equilibrio, mostran-

$\max\{d(x,y), d(y,z)\}$.

[12] Nel presente scenario il parametro d'ordine è q_{ab} mentre il parametro di controllo è la temperatura T che regola l'intensità con cui gli spin percepiscono l'accoppiamento: di fatto vedremo che concettualmente è lo stesso parametro di controllo della mappa logistica.

[13] Questa proprietà emegente si renderà particolarmente interessante nello studio di un particolare vetro di spin: la rete neurale. Si veda a tal proposito il nostro contributo sul volume di Ithaca dedicato all'Intelligenza Artificiale.

do fenomeni non canonici in Fisica, quali l'*aging*. Infatti, questi vetri invecchiano[14] rompendo uno dei capisaldi della Fisica, l'invarianza traslazionale nel tempo degli esperimenti (il principio di Galileo). Per chiarire questo concetto, assumiamo di fare un esperimento su un vetro di spin, per esempio di misurarne la suscettività in risposta ad un campo magnetico, ed assumiamo che l'esperimento debba durare 24 ore: troveremo che, se facciamo l'esperimento in momenti differenti della vita del vetro di spin avremo risposte differenti[15], alla stregua di quanto ci aspettiamo da un sistema biologico, infatti, citando Peter Sollich: *One of the core ideas of statistical mechanics is that equilibrium states can be accurately described in terms of only a small number of thermodynamic variables, such as temperature and pressure. For glassy systems, which can remain far from equilibrium on very long time scales, no similar simplification exists a priori; the whole past history of a sample is in principle required to specify its state at a given time* [7].
In relazione alla farraginosità della Matematica della Meccanica Statistica complessa, si pensi che nella prima formulazione – mediante il *replica trick* – il conto prevede di esprimere un logaritmo mediante la relazione $N^{-1} \ln Z = \lim_{n\to 0} \frac{Z^n - 1}{nN}$, in cui N rappresenta il numero di spin ed n quello delle repliche: poiché il numero di coppie di repliche che possono supplire informazione a q_{ab} è $n(n-1)/2$, prendendo un qualunque $n \in (0, 1)$, alla volta del prolungamento analitico $n \to 0$, il numero di coppie con cui si ha a che fare è negativo... Michel Talagrand, uno dei più stimati probabilisti francesi (e che ha giocato un ruolo cruciale nel riformulare in maniera matematica rigorosa la teoria dei vetri di spin), ha chiamato il suo libro *Spin Glasses: a challenge for Mathematicians* [5].

Parte Tre (dinamica): Semplice versus Complesso nei Sistemi Dinamici

Per introdurre il lettore al chaos deterministico (e.g., l'*effetto farfalla*) di solito si usa il modello di Edward Lorenz di impiego nella metereologia, ma qui, prefe-

[14] Si può anche dimostrare rigorosamente che ogni step di rottura di simmetria di replica (RSB) aggiunge una scala di tempo alla termalizzazione ed, essendo la soluzione del modello di SK *full-RSB*, la termalizzazione di fatto non si completa mai.

[15] Il fenomeno dell'aging implica anche la violazione da parte di questi sistemi del Teorema di Fluttuazione-Dissipazione. L'aging si può apprezzare anche nei vetri strutturali per esempio osservando come questi filtrano la luce in vecchie abbazie che li espongono secolari: l'intensità della luce nella parte alta del vetro è maggiore perchè il vetro *sta colando* su scala di tempo dei millenni per effetto della forza di gravità.

rendo la Biologia alla Fisica, discuteremo il modello del biologo Robert May che descrive la crescita di una cultura batterica, e vedremo come la caoticità mostri profonde similitudini con la complessità.
Per onestà intellettuale è d'obbligo notare che il padre ante litteram del chaos deterministico è, a buona ragione, Henry Poincarè: infatti nei suoi piccoli denominatori, che trovava per via perturbativa studiando il problema dei tre corpi, aveva già di fatto inferto il colpo letale al determinismo, anche in sistemi del tutto scevri di una trattazione probabilistica. Questa materia è oggetto di studio dei *Sistemi Dinamici* più che della *Meccanica Statistica*, sebbene le due discipline siano fortemente legate, perennemente congiunte dalla Teoria Ergodica, la loro fede nuziale.

Il chaos deterministico

Consideriamo una concentrazione batterica libera di duplicarsi in un disco di Petri[16], ed assumiamo che sia $C = 0$ la concentrazione nulla e $C = 1$ la concentrazione massima (per la quale l'intero disco è saturo di batteri), quindi $C \in (0, 1)$. Assumiamo inoltre che, come certamente ragionevole all'inizio (per concentrazioni basse, dove non ci sia necessità di competere per le risorse) un buon modello per gestire la loro proliferazione sia l'ormai classico sistema dinamico $\dot{C} = AC$, ma in una sua versione discretizzata e, al fine di tenere la trattazione originale (chiamando il parametro $A \to r$), scriviamo la mappa

$$C_{t+1} = rC_t,$$

dove C_t rappresenta la concentrazione al passo iterativo t-esimo. L'equilibrio, la soluzione per tempi lunghi di questo modello, è trasparente: se $r > 1$ i batteri colonizzano l'intera cella di Petri, se $r < 1$ muoiono e la cella rimane vuota: due sole soluzioni limite, proprio come nel Curie-Weiss dove gli spin possono interagire solo in maniera coesiva e non competono.
Ma cosa succede se vogliamo rendere il modello più realistico ed in prima istanza ci rendiamo conto che questi batteri non vivono nel giardino dell'Eden ma nel disco del Petri in cui il loro nutrimento esiste in quantità finita? La mappa precedente deve essere rivista per tener conto di effetti competitivi: per questo, nel 1976, Robert May introdusse la *mappa logistica* che costituisce una versione discreta del modello demografico precedentemente introdotto dal matematico Pierre François Verhulst nel quale siano presenti sia

[16] La piastra di Petri è un recipiente piatto di vetro o plastica, solitamente di forma cilindrica e costituisce un importante strumento di lavoro in molti campi della Biologia, in particolare per la crescita di colture cellulari.

coesione che competizione, cioè

$$C_{t+1} = rC_t \cdot (1 - C_t), \tag{7}$$

il cui significato è di nuovo trasparente – il fattore $1 - C_t$ determina un rallentamento nella crescita della concentrazione, tanto più significativo quando più grande è la concentrazione – ma la sua soluzione stavolta meno. Come sintetizzato in figura 6 se $r \in [0, 1)$, la popolazione calerà fino a morire, indipendentemente dal valore iniziale della popolazione;
se $r \in [1, 3)$, la popolazione andrà a stabilirsi al valore $(r - 1)/r$;
se $r \in [3, r_c)$, per quasi tutte le condizioni iniziali, la popolazione arriva ad oscillare indefinitamente tra un certo numero (crescente con r e potenza di 2) di valori dipendenti da r, si ha cioè una cascata di biforcazioni con raddoppiamento del periodo;
se $r > r_c \approx 3.56995$ si ha l'insorgenza del caos.

Quando il sistema si trova in una fase caotica, anche se conosciamo l'equazione che governa il modello e la condizione di Cauchy, è comunque impossibile prevedere cosa succederà per tempi lunghi. Cerchiamo di approfondire il concetto: per sviscerare meglio la genesi dell'erraticità della mappa logistica, analizziamola per un particolare valore del parametro di controllo, per $r = 4$ (in pieno regime caotico quindi). Con $r = 4$ esiste un cambio di variabili fortuito che rende trasparenti le iterazioni della mappa logistica: $C_t = [1 - \cos(2\pi\theta_t)]/2$. Passando a variabili angolari l'equazione (7) diventa $\theta_{t+1} = 2\theta_t$.
Scriviamo ora la condizione iniziale θ_0 in un alfabeto binario per semplicità, per esempio $\theta_0 = (0, 1, 1, 1, 0, 1, 0, 1)$ cioè $0 + 1/2 + 1/8 + 1/16 + 0/32 + 1/64 + 0/128 + 1/256 = (128 + 64 + 32 + 8 + 2)/256 = 117/128 \sim 0.9$. Prima di procedere, senza scomodare Werner Heisenberg, ricordiamo che qualunque strumento di misura ha una risoluzione finita, quindi è impossibile conoscere perfettamente la condizione iniziale (la stringa di numeri binari di cui sopra deve essere finita). Quest'osservazione è cruciale perchè la mappa logistica in alfabeto binario, per $r = 4$, non fa altro che prendere la stringa iniziale e, ad ogni iterazione, traslarla rigidamente *verso sinistra, di un'unità*, quindi $\theta_1 = (1, 1, 1, 0, 1, 0, 1, ?) \to \theta_2 = (1, 1, 0, 1, 0, 1, ?, ?)$, etc... Pertanto, se conosciamo la condizione iniziale con N cifre significative, dopo N interazioni la mappa avrà perso completamente ricordo della condizione iniziale θ_0 e si comporterà come un cammino aleatorio (e.g., con uno spettro di potenza bianco)[17].
Questa caratteristica della mappa logistica, un sistema dinamico non-lineare, è oltremodo diversa rispetto alla sua controparte lineare, infatti è proprio nella

[17] Rimandiamo alla digressione sulla temperatura ubriaca affrontata nel box.

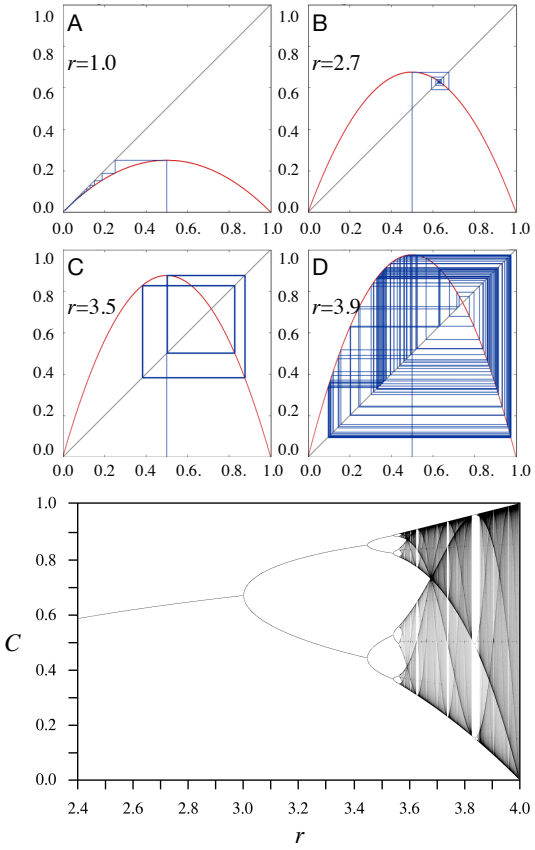


Figura 6: *Diagramma a ragnatela o diagramma di Verhulst (in alto) e diagramma di biforcazione (in basso) per la mappa logistica (7). Attraverso il diagramma a ragnatela è possibile dedurre lo stato a lungo termine di una condizione iniziale in seguito all'applicazione ripetuta della mappa: un punto fisso stabile corrisponde ad una spirale interna (A, B), un punto fisso instabile corrisponde ad una esterna, un'orbita di periodo 2 è rappresentata da un rettangolo (C), mentre cicli di periodo maggiore producono linee chiuse di forma più complessa, un orbita caotica apparirebbe invece come un'area densamente colorata (D) ad indicare un numero infinito si valori non ripetuti. Nel diagramma di biforcazione l'asse orizzontale mostra i valori del parametro r, mentre quello verticale mostra il relativo valore di C_t, con t che tende all'infinito.*

linearità che lo schema di riduzione affonda le sue radici: l'oscillatore armonico con cui abbiamo iniziato, a dire $\frac{d^2y}{dt^2} = -\omega^2 y$, o scritto in termini di sistema dinamico $\dot{x} = Ax$ come

$$\frac{dx_1}{dt} = x_2, \quad \frac{dx_2}{dt} = -\omega^2 x_1,$$

ha soluzione esplicita

$$\begin{aligned} x_1(t) &= x_1(0)\cos(\omega t) + x_2(0)\omega^{-1}\sin(\omega t), \\ x_2(t) &= x_2(0)\cos(\omega t) - x_1(0)\omega^{+1}\sin(\omega t), \end{aligned}$$

che porta ad una stima dell'incertezza nella sua evoluzione sempre limitata e pari a

$$\begin{aligned} |\delta x|^2 &= \delta x_1^2(t) + \delta x_2^2(t) \\ &< (1 + \omega^{+2} + \omega^{-2})(\delta x_1^2(0) + \delta x_2^2(0)), \end{aligned}$$

d'altronde, per i sistemi lineari, l'incertezza rimane sempre polinomiale, a dire $|\delta x(t)| := |x(t) - x'(t)| \sim \epsilon(1 + \mathcal{P}(t))$ con $\mathcal{P}(t)$ non esponenziale ($\mathcal{P}$ sta per "polinomiale") [6].

Nel vedere il paradigma classico (storicamente egemone) di un mondo governato da leggi deterministiche (e, in un certo senso, rassicurantemente prevedibile) sgretolarsi non appena si abbandona il telaio del (vero) oscillatore armonico e nel veder nascere spontaneamente l'ordine dal caso, ci si ferma a riflettere (d'altronde che una costante di Liptshitz positiva nell'evoluzione delle soluzioni di un'equazione differenziale ci avvertisse di una potenziale dipartita esponenziale tra due traiettorie inizialmente contigue era noto per fino a Liptshitz...).
Forse con il determinismo del riduzionismo abbiamo costruito un'Ingegneria che ci permette di leggere di notte e volare da un continente all'altro, ma -ad oggimeno serve a comprendere fenomeni biologici, dall'omeostasi della singola cellula alla costruzione della Biblioteca di Alessandria (non nell'accezione tecnica del sostantivo *costruzione*, quello lo sappiamo fare, ma come esigenza collettiva emergente della nostra necessita' di sapere), di contro, proprio nel comprendere i limiti dell'approccio di riduzione troviamo una via per superarlo: è una via la cui pavimentazione è iniziata da poco se si guarda al progresso scientifico nel suo complesso, ma i suoi lastroni vengono posizionati con un rinato positivismo (di cui, a nostro avviso, un esempio calzante è offerto dagli sforzi di Luca Peliti [8] nel riformulare mediante la meccanica statistica degli spin-glasses la teoria neutrale dell'evoluzione di Mooto Kimura [9]).
In concerto con il contributo dei colleghi Nando Boero e Giampaolo Co' presente sul numero parallelo di Ithaca Educational dedicato ai Sistemi Complessi, per chiudere questo articolo divulgativo parafrasando un padre fondatore che ha contribuito tanto al telaio riduzionista quanto lo ha messo in dubbio ponendo i pilastri della Teoria della Probabilità, *ce que nous connaissons est peu de chose, ce que nous ignorons est immense.*

[1] D.J. Amit: *Modeling brain function* Cambridge Press, Cambridge (1985).

[2] S.N. Durlauf: *How can statistical mechanics contribute to social science?*, Proceedings of the national academy of sciences, 96 (1999) 10582.

[3] G. Parisi, M. Mezard, M.A. Virasoro: *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications* World Scientific Publishing, Singapore (1987).

[4] F. Guerra: *Broken replica symmetry bounds in the mean field spin glass model*, Communications in Mathematical Physics, 233 (2003) 1.

[5] M. Talagrand: *Spin glasses: a challenge for mathematicians* Springer Science, Paris (2003).

[6] A. Vulpiani, et al.: *Chaos and coarse graining in statistical mechanics* Cambridge Press, Cambridge (2008).

[7] P. Sollich: *Fluctuation-dissipation relations and effective temperatures in simple non-mean field systems*, Journal of Physics C, 12 (2002) 1683.

[8] L. Peliti: *Introduction to the statistical theory of Darwinian evolution* Lectures at the Summer College on Frustrated System, Trieste (1997).

[9] M. Kimura: *The neutral theory of molecular evolution* Cambridge University Press, Cambridge (1983).

Elena Agliari: è ricercatrice in Fisica Matematica presso Sapienza Università di Roma, dove insegna -tra i vari- *Metodi Matematici per le Reti Neurali*. Si occupa di principalmente di Meccanica Statistica dei Sistemi Complessi, Teoria dei Grafi e Processi Stocastici, con particolare attenzione alle loro applicazioni nella Biologia e nell'Intelligenza Artificiale. Ha all'attivo un centinaio di articoli scientifici di cui gran parte coautorati con Adriano Barra e Francesco Guerra.

Adriano Barra: è professore associato in Fisica Matematica presso l'Università del Salento, dove insegna -tra gli altri- *Metodi Matematici per l'Intelligenza Artificiale* (ed in passato *Sistemi Complessi* per la Scuola Superiore ISUFI). Si occupa di principalmente di Meccanica Statistica dei Sistemi Complessi, Teoria dei Grafi e Processi Stocastici, con particolare attenzione alle loro applicazioni nella Biologia e nells'Intelligenza Artificiale. Ha all'attivo un centinaio di articoli scientifici di cui gran parte coautorati con Elena Agliari e Francesco Guerra, di cui è stato il collaboratore primario per quindici anni (dal 2002 al 2017).

Numero XVI Anno 2020

Viaggio nella Scienza

Ithaca

Intelligenza artificiale