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

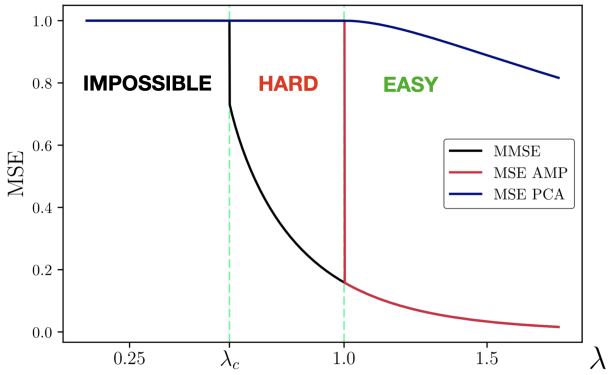

Figura 1: *Ita: Da [20]. Grafico dello spike-MMSE, il MSE dell'algoritmo AMP e della PCA naive per il modello SW con prior* (15) *con* $\rho = 0,05$. *Si osservano transizioni di fase del primo ordine sia di teoria dell'informazione che algoritmica. È presente un gap computazionale-statistico (fase difficile) 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].

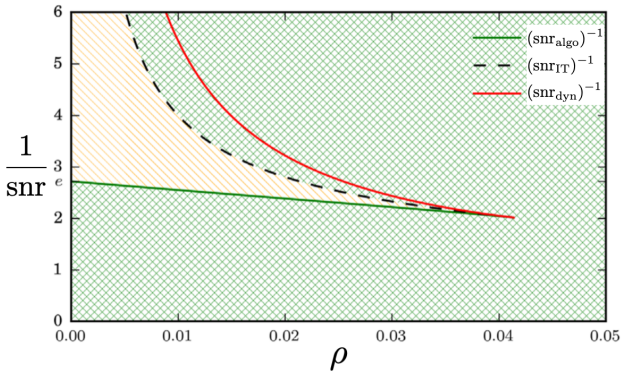

Figura 2: *Ita: Da [21]. Diagramma di fase del modello SW con parametri di Bernoulli $x_i \sim Ber(\rho)$ in funzione della scarsità ρ e del reciproco del RSR totale dato da RSR $:= \lambda\rho^2$. Non c'è transizione di fase nel sistema se $\rho > 0,0414$ e una transizione di fase del primo ordine altrimenti. 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.

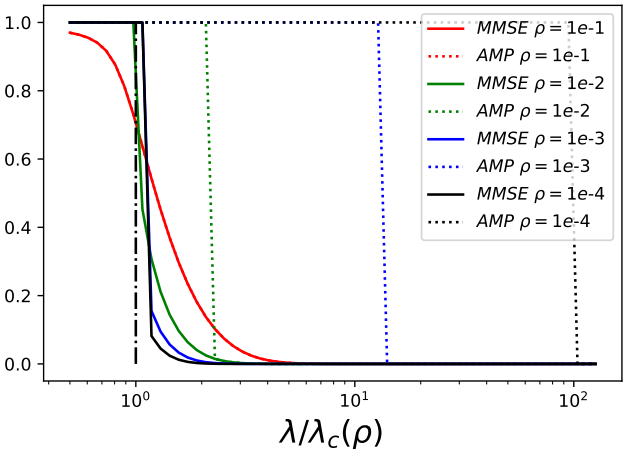

Figura 3: *Ita:Da [19]. Transizioni di fase in teoria dell'informazione del tipo tutto-o-niente e transizioni di fase algoritmiche ottenute mediante l'AMP per il modello SW con parametri di Bernoulli $x_i \sim Ber(\rho)$. Man mano che la scarsità ρ descresce, entrambe le transizioni diventano più nitide: 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.