Elements of statistical inference and statistical physics

miguel.berganza@roma1.infn.it

Contents

1 The direct problem: examples **2**

1.1 Sorting random numbers according to a given single-valued distribution 2

1.2 Sorting random vectors according to a multi-dimensional distribution: the Monte-Carlo method 2

1.2.1 Markov-Chain Monte Carlo 4

1.2.2 Gibbs sampling (heatbath) algorithm. 5

1.3 Variational free energy approximation of an interacting model 5

2 Inference. The inverse problem in examples **8**

2.1 Bayesian estimators 8

2.2 Maximum likelihood infering a Gaussian distribution 8

2.3 Inferring a mixture of Gaussian distributions 9

2.4 Maximum likelihood inferring the coupling parameters of a statistical model. . . . 10

2.4.1 Inference of a model with real degrees of freedom (from pairwise correlations) 10

2.4.2 Inference of Ising degrees of freedom in the linear response approximation (from pairwise correlations) 11

3 Maximum entropy inference **13**

3.1 General formulation 13

3.2 Maximum entropy inference with pairwise correlations 14

3.2.1 Applications of maximum entropy inference 15

4 Elements of Bayesian model selection **17**

5 Notions of probabilistic models of cognition **19**

5.1 A simple illustration of model selection (coin flipping) 19

5.2 Inferring from two cues. 20

5.3 A mention to the Free Energy Principle 21

6 Learning in Neural networks **22**

6.1 Learning in Restricted Boltzmann Machines 22

6.2 Two examples of unsupervised learning in RBM's 24

A Appendix: the Metropolis algorithm **26**

A.1 A Metropolis algorithm for the likelihood estimation of the Gaussian mixture inference problem 26

B Appendix: cumulant expansion **27**

B.1 One-dimensional cumulant expansion 27

B.2 Multivariate cumulant expansion 27

C The Gaussian Model and the Wick's theorem **28**

Abstract

Compressed lecture notes of a six-hour seminar regarding unsupervised statistical inference and probabilistic models of cognition. We will briefly describe some notions in statistical inference (maximum likelihood inference, difference among correlations and effective interactions, over-fitting and Bayesian model selection) from a statistical-physics perspective. Afterwards, some aspects of probabilistic (Bayesian) models of cognition will be briefly mentioned. The

aim of the lessons is not to provide an exhaustive nor self-contained view of these arguments, but to introduce graduation students in physics to some basic concepts, drawing analogies with concepts from their background in statistical mechanics.

1 The direct problem: examples

By *direct problem* we mean to generate (to *sample*) a finite number of *particular* instances, according to a given, *general* rule.

1.1 Sorting random numbers according to a given single-valued distribution

A simple particular case is that of extracting a finite set of real numbers, $\{x^{(s)}\}_{s=1}^{n}$ from a known probability distribution f, $\int_{\mathbb{R}} f = 1$.

Cumulative method. To sample a probability distribution f, of which we know its (invertible) primitive function, F, one samples ξ, uniformly distributed in $[0, 1]$, and returns $F^{-1}(\xi)$. Check (if you want) that the resulting number is, indeed, distributed according to f.

***Exercise 1.** *What about a distribution of which one does not known the analytical form for its cumulative, as the Normal distribution? It is still possible to use the cumulative method (Box-Mueller algorithm, 1985). The algorithm is: 1) one generates two numbers, u. d. in* $[0, 1]$*,* ξ_r*,* ξ_θ*; 2) the couple of normally distributed random numbers is:* $X = (-2\sigma^2 \ln(1 - \xi_r)) \cos(\xi_\theta)$ *and* $Y = (-2\sigma^2 \ln(1 - \xi_r)) \sin(\xi_\theta)$*. Check that the numbers* X, Y *so generated are, indeed, distributed according to a* $(0, \sigma)$*-normal distribution,* $\mathcal{N}(\cdot|0, \sigma)$*.*[1]

Exercise 2. *Develop and numerically check the Box-Mueller algorithm.*

Crude MC integration. To integrate a one dimensional function f in $[a, b]$, one: generates ξ, u. d. in $[0, 1]$; computes $x \equiv a + (b - a)\xi$ and $f_j \equiv f(x)$; one repeats these two operations for $j = 1, \ldots, n$. Afterwards, the integral of f, I, can be estimated as $I \simeq (b - a)\langle f\rangle_n$ where $\langle f\rangle_n = \sum_{j=1}^{n} f_j/n$. In particular, f_j are random numbers distributed with average $\langle f\rangle_\infty = I/(b - a)$ and variance $\sigma^2 = \langle f^2\rangle_\infty - \langle f\rangle_\infty^2$, the central limit theorem wants the variable $w_n = n^{1/2}(\langle f\rangle_n - \langle f\rangle_\infty)$ to be distributed $(0, \sigma)$-normally for large n, in other words:

$$\langle f\rangle_n = \langle f\rangle_\infty + n^{-1/2}\sigma y \tag{1.1}$$

where y is a standard Gaussian variable. The generalisation to d-dimensional functions is straightforward, the variance of the estimation is $\sim n^{-1}$, independent of d.

1.2 Sorting random vectors according to a multi-dimensional distribution: the Monte-Carlo method

A more complex case is that of sorting a finite set of multi-dimensional vectors, $\{\boldsymbol{\sigma}^{(s)}\}_{s=1}^{n}$ from a known multi-dimensional probability distribution $\pi(\boldsymbol{\sigma})$.

Consider a discrete space, $\mathcal{N}$-dimensional space $\boldsymbol{\Sigma}$ of *configurations*, $\boldsymbol{\Sigma} = \{\boldsymbol{\sigma}_\alpha\}$, $\alpha = 1, \ldots, \mathcal{N}$.[2] Consider a target probability distribution $\boldsymbol{\pi} = (\pi_\alpha)_{\alpha=1}^{\mathcal{N}}$, $\pi_\alpha = \text{prob}(\boldsymbol{\sigma}_\alpha)$ for $\alpha = 1, \ldots, \mathcal{N}$. The probabilities π_α are non-factorisable in the sense that they cannot be expressed as a product of the various components (or particles) of the configuration: $\pi_\alpha \neq \prod_i p_i(\sigma_{\alpha,i})$, where p_i is the probability distribution of the single degree of freedom (think, for example, in a Boltzmann probability distribution in the canonical ensemble, with energy given by a pairwise interaction Hamiltonian).

[1] Mind the Gaussian integral equation: $\int e^{-Ax^2+Bx} = (\pi/A)^{1/2} e^{B^2/4A}$.

[2] The set of configurations may be concieved as the phase space of a physical system. Each state $\boldsymbol{\sigma}_\alpha$ of the phase space is composed by N components (or particles, in physical language), $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1^{\otimes N}$, where $\boldsymbol{\Sigma}_1$ is the D-dimensional single-particle space (or *degree of freedom*). For instance, if the single-particle degree of freedom are Ising, binary spins, $\boldsymbol{\Sigma}_1 = \{0, 1\}$, $D = 2$ and $\mathcal{N} = D^N$.

Example 1. *Use of configuration sampling in Statistical Physics. Why should one be interested in sorting configurations according to their Boltzmann (or the probability corresponding to a different thermodynamic ensemble, e.g., macro-canonical) probability distribution, P? The answer is: to solve the direct problem in physics, i.e., to estimate efficiently (in polynomial time) the ensemble-average of observables $O : \mathbf{\Sigma} \to \mathbb{R}$, $\langle O \rangle$, where $\langle \cdot \rangle = \sum_\alpha \cdot P(\boldsymbol{\sigma}_\alpha)$ is the ensemble average. If one is able to sort T configurations $\boldsymbol{\sigma}(t)$, $t = 1, \ldots, T$ from P, it is possible to provide an approximation for $\langle O \rangle$, whose error decreases as $T^{-1/2}$. Indeed, using crude MC integration, the estimation $\langle O \rangle_T = \frac{1}{T}\sum_{t=1}^{T} O(\boldsymbol{\sigma}(t))$ is expected to behave as the stochastic variable:*

$$\langle O \rangle_T = \langle O \rangle + y \frac{s}{T^{1/2}} \tag{1.2}$$

where y is a standard Gaussian variable, $y \sim \mathcal{N}(0,1)$ and $s^2 = \langle O^2 \rangle - \langle O \rangle^2$ is the variance of O according to P.

Exercise 3. *According to the last equation, and since $s \sim N^{-1/2}$, one could think that it is more efficient to sample configurations of systems with a large number N of particles, in order to obtain a more accurate estimation. In fact, there are at least three reasons why this is not this way. Can you guess some (hint: see footnote 4, in particular "Sources of correlations in phase transitions" in the mentioned notes)?*

If the probability distribution is non-factorisable (as the Boltzmann probability distribution of a system of interacting particles, in a statistical physical language), the sorting of $\boldsymbol{\sigma}$ from P cannot be done *uniformly*, i.e., sampling configurations $\boldsymbol{\sigma}$ in which each component σ_i is extracted independently with a given probability distribution p_i. Uniform sampling is not efficient for the numerical estimation of observables.

Example 2. *Inefficiency of uniform Monte Carlo (MC) sampling in the canonical ensemble. Recall the canonical ensemble: at inverse temperature β, one is interested in a probability distribution for ϵ, the intensive energy, given by $p_\beta(\epsilon) = \exp[-N\beta\tilde{\phi}(\beta,\epsilon)]/Z_\beta$, where $\tilde{\phi} = (\epsilon - Ts)$ is the intensive free energy functional (and s is the microcanonical entropy), N is the system mass, and Z is the partition function. In saddle-point approximation,*[3] *it is $Z_\beta = \exp[-N\beta\phi(\beta)]$, where $\phi(\beta) = \tilde{\phi}(\beta, \epsilon_\beta) = \min_\epsilon \tilde{\phi}(\beta,\epsilon)$ is the intensive free energy. The probability of finding a configuration with energy ϵ', different from the most probable energy ϵ_β, is, hence, $p_\beta(\epsilon') = \exp[-N\beta(\tilde{\phi}(\beta,\epsilon') - \tilde{\phi}(\beta,\epsilon_\beta))]$, which is exponentially suppressed in N. It follows that random configurations (as those sampled by uniform MC), that correspond to $\beta = 0$, have intensive energy $\epsilon_{\beta=0}$ with very high probability. Vice-versa, they exhibit vanishing (exponentially small in N) probability in an ensemble at $\beta > 0$.*

***Exercise 4.** *Convince yourself of $p_\beta(\epsilon') = \exp[-N\beta(\tilde{\phi}(\beta,\epsilon') - \tilde{\phi}(\beta,\epsilon_\beta))]$ in the saddle-point approximation (hence, of the equivalence of the micro-canonical and canonical ensembles).*

***Exercise 5.** *The central limit theorem (see, v.g., [Marinari and Parisi, 2004]) states that if $z^{(i)}$ are n independent, random variables distributed with a distribution exhibiting average and standard deviation μ, $\sigma < \infty$ respectively, then the probability distribution of their empirical average $\langle z \rangle_n$ converges, for large n, to the normal distribution $\mathcal{N}(\mu, \sigma/n^{1/2})$. Convince yourself of equation (1.2), in the light of the central limit theorem, using the identification $z^{(i)} = O(\mathbf{x}^{(i)})$, where the $\mathbf{x}^{(i)}$ are sampled from the target P (and, hence, $O(\mathbf{x}^{(i)})$ is a stochastic variable with average $\langle O \rangle$ and variance $\langle O^2 \rangle - \langle O \rangle^2$).*

Exercise 6. *If one samples instead $\mathbf{x}^{(i)}$ uniformly, the estimator $\bar{O} = (1/n)\sum_{i=1}^{n} P(\mathbf{x}^{(i)})O(\mathbf{x}^{(i)})$ does not converge for large n to $\langle O \rangle$, since the variable $P(\mathbf{x}^{(i)})O(\mathbf{x}^{(i)})$ exhibits null average when the $\mathbf{x}^{(i)}$ are sampled uniformly (please, check). If, instead, one generates n configurations with uniform MC sampling, and constructs the sum $\sum_{i=1}^{n} P(\mathbf{x}^{(i)})O(\mathbf{x}^{(i)})$, the terms in the sum will be exponentially small in N (if P is not factorisable, i.e., if $\beta > 0$, see Example 2). Hence, for an n that grows polynomically in N, the sum does not provide a useful estimator of $\langle O \rangle$.*

[3] For $N \to \infty$, $s, f : \mathbb{R}^n \to \mathbb{R}$ it is

$$\int \mathrm{d}\mathbf{x}\, f(\mathbf{x}) e^{Ns(\mathbf{x})} = (2\pi/N)^{n/2} \sum_k e^{Ns(\mathbf{x}^{(k)})} f(\mathbf{x}^{(k)})\, [\det - \mathrm{Hes}[s](\mathbf{x}^{(k)})]^{-1/2} \tag{1.3}$$

where $\mathbf{x}^{(k)}$ are the saddle points of s.

1.2.1 Markov-Chain Monte Carlo

Markov Chains. A Markov Chain is a probability measure over a sequence of configurations, such that the conditional probability of having $\boldsymbol{\sigma}^{(t)}$, the configuration at time t depends only on $\boldsymbol{\sigma}^{(t)}$, $\boldsymbol{\sigma}^{(t-1)}$. The transition probabilities can be cast into a matrix T whose element $T_{\alpha\beta}$ is the transition probability of the α-th to the β-th state, $\alpha, \beta = 1, \dots, \mathcal{N}$. The transition matrix is a stochastic matrix, it satisfies: $T_{\alpha\beta} \geq 0 \,\forall \alpha, \beta$ and $\sum_\beta T_{\alpha\beta} = 1$. The Markov Chain characterized by T is said *irreducible* if given any two states α, β, one can reach β form α in a finite time, i.e., if $\forall\alpha, \beta$ there exists n such that $(T^n)_{\alpha\beta} > 0$. A stronger property is *aperiodicity*: if there exists a n such that $(T^t)_{\alpha\beta} > 0$ for all α, β, and for all $t > n$.

The matrix T along with the probability distribution for the first element of the chain, $\pi^{(0)}$, define the Markov Chain, and induce a probability measure on the set of n sequences of states, $\boldsymbol{\sigma}_{\alpha_1}, \boldsymbol{\sigma}_{\alpha_2} \dots$, which is $\pi^{(0)}(\boldsymbol{\sigma}_{\alpha_1})\, T_{\alpha_1\alpha_2} T_{\alpha_2\alpha_3} \dots$, and the probability of having the β-th state at time t is $= \sum_\alpha (T^t)_{\alpha\beta} \pi^{(0)}(\boldsymbol{\sigma}_\alpha)$.

Theorem 1. *Discrete, aperiodic, irreducible Markov Chains are such that*

1. *The limit $\pi_\beta = \lim_{n\to\infty}(T^n)_{\alpha\beta}$ uniquely exists, independently on α. $\pi_\beta \equiv \pi(\boldsymbol{\sigma}_\beta)$ is a PD ($\sum_\beta \pi_\beta = 1$), stationary under T:*

$$\pi_\beta = \sum_\alpha T_{\alpha\beta} \pi_\alpha \qquad \text{Balance condition} \tag{1.4}$$

2. *If $f \in l^2(\pi)$ (square-integrable with respect to π) and $f_\alpha \equiv f(\boldsymbol{\sigma}_\alpha)$, it is:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n} f(\boldsymbol{\sigma}^{(t)}) = \sum_{\alpha=1}^{\mathcal{N}} \pi_\alpha f_\alpha \tag{1.5}$$

regardless of $\pi^{(0)}$, the fluctuations for finite n being of order $n^{-1/2}$.

The dynamic (or Markov-Chain) Monte Carlo method consists in choosing a transition matrix T such that its stationary distribution $\boldsymbol{\pi}$ is the desired one. The theorem before requires for the dynamic MC method to work, that 1) T must be irreducible and 2) that it satisfies the Balance condition. In these circumstances, one can iterate T starting from an arbitrary configuration and, after a sufficiently high number of iterations, obtain as much as desired configurations sampled from $\boldsymbol{\pi}$, and the ensemble averages of observables according to $\boldsymbol{\pi}$.[4]

We notice that equation (1.2) is true only if the configurations $\boldsymbol{\sigma}(t)$ are sampled independently in the Markov Chain. If they are correlated in t (if consecutive observables $O(\boldsymbol{\sigma}(t))$ at nearby values of t tend to be similar), there is a further factor proportional to y, although the dependence in T is still $T^{-1/2}$. The proportionality factor is $\tau_{\rm i}^{1/2}$:

$$\langle O \rangle_T = \langle O \rangle + y \frac{s}{(T/\tau_{\rm i})^{1/2}}, \tag{1.6}$$

$\tau_{\rm i}$ is the so called *integrated* or *correlation* time, associated to the observable O and to the way in which the configurations are sorted from P (i.e., to the specific Markov-Chain Monte-Carlo algorithm). Roughly speaking, $\tau_{\rm i}$ is such that values of O computed for configurations that are distant more than $\tau_{\rm i}$ times in the realisations of the Markov chain sequence, are typically uncorrelated. In this sense, eq. (1.6) for a correlated sequence is like using eq. (1.2) but measuring the time in units of $\tau_{\rm i}$.

***Exercise 7.** *Convince yourself that, if $(T^m)_{\alpha\beta}$ may be taken as equal to π_β for sufficiently large m, then $\boldsymbol{\pi}$ is stationary under T (the Balance Condition).*

Exercise 8. *Let us define the distance between two distributions $||\boldsymbol{A} - \boldsymbol{B}|| = \sum_\beta |A_\beta - B_\beta|$. Consider a Markov Chain satisfying the balance condition, and show that the distance between a vector $\boldsymbol{v}$ and the stationary distribution $\boldsymbol{\pi}$ is larger than that between $\boldsymbol{v}^\dagger T$ and $\boldsymbol{\pi}^\dagger$ (use the triangle inequality).*

[4] See some the limitations of the MC method in [Ibanez-Berganza, 2016], and references therein.

Exercise 9. *The Perron-Frobenius theorem states that the largest absolute value of the eigenvalues of an irreducible Markov matrix T is 1, and that there is at least an eigenvalue $\lambda_1 = 1$ (whose corresponding eigenvector is the stationary probability distribution). Now, suppose that the second largest (in modulus) eigenvalue of T, λ_2, is such that $|\lambda_2| < 1$. Then, demonstrate that, after m steps of the Markov chain, the convergence to the stationary distribution is exponentially fast in m, with a characteristic time equal to $\tau_{\mathrm{e}} = -1/\ln|\lambda_2|$. τ_{e} is called exponential time of the Markov Chain.*

1.2.2 Gibbs sampling (heatbath) algorithm.

We define the transition matrix of the *heat bath* algorithm as $p^{(m)}[\boldsymbol{\sigma} \to \boldsymbol{\sigma}'] = \pi^{(m)}(\sigma'^{(m)}|\boldsymbol{\sigma}_{\setminus m})$, equal to the marginal stationary probability distribution of the m-th particle degree of freedom, given the rest of the configuration $\boldsymbol{\sigma}_{\setminus m}$, and new and old configurations being equal except by the m-th particle, $\boldsymbol{\sigma}'_{\setminus m} = \boldsymbol{\sigma}_{\setminus m}$. In other words, the *Gibbs sampling* or *heatbath* algorithm proposes a new state of particle m with its marginal stationary probability, independently of the current state of particle m. Many particles can then be sequentially or randomly updated. The *MC sweep* or the (sequential or random) sequence of N Gibbs MC transitions for different particles results to satisfy balance, and is aperiodic (for a demonstration of these properties see [Fischer and Igel, 2012]).

In figure 1.1 we present an illustration of the Monte Carlo method: one has applied the Gibbs sampling algorithm to the canonical ensemble sampling of the Potts model on the 2D lattice. For each sample, the energy of the sampled configuration is plotted as a function of the number of MC *sweeps* (a MC *sweep* is a sequence of N sequential or random transitions). Various updating algorithms are compared: the single-spin Gibbs sampling algorithm with sequential updates, the Metropolis algorithm (see sec. A) with sequential and random updates, and the two-hit Metropolis (multi-hit means to perform multiple successive MC updatings of a single particle). The initial conditions in different runs are either completely ordered or disordered configurations.

Exercise 10. *Consider the Gibbs sampling algorithm (with random updating) for the sampling of the Ising model on an arbitrary lattice*[5]*, in the canonical ensemble at inverse temperature β. The transition probability among two configurations differing by the m-th spin ($m = 1, \dots, N$) is given by the marginal probability $p^{(m)}[\boldsymbol{\sigma} \to \boldsymbol{\sigma}'] = \pi^{(m)}(\sigma'^{(m)}|\boldsymbol{\sigma}_{\setminus m})$. Write such marginal probability distribution. Can this transition matrix be parallelized by chosing two different random spins, r, s, and updating them simultaneously with probability $p^{(r)}[\boldsymbol{\sigma} \to \boldsymbol{\sigma}']\, p^{(s)}[\boldsymbol{\sigma} \to \boldsymbol{\sigma}']$? Under what conditions is this a valid MC method?*

Exercise 11. *Why in fig. 1.1 the Gibbs sampling algorithm reaches the equilibrium before (after a lower number of MC sweeps) than the Metropolis algorithm? Is this a general fact? Does make any sense to perform multi-hit Gibbs sampling?*

1.3 Variational free energy approximation of an interacting model

We consider probability distribution on the set of many-particle configurations $\mathbf{x} = \{x_i\}_{i=1}^{n}$ where x_i is the i-th degree of freedom. The probability distribution is given by the energy functional $E[\mathbf{x}, J]$, depending on the set of parameters J:

$$\mathcal{L}(\mathbf{x}|\beta, J) = Z(\beta, J)^{-1} \exp\left[-\beta E[\mathbf{x}, J]\right] \tag{1.8}$$

$$\tag{1.9}$$

where Z is the partition function, normalizing $\mathcal{L}$. One desires to approximate $\mathcal{L}$ by a probability distribution $Q(\mathbf{x}; \boldsymbol{\theta})$ on a set of variational parameters $\boldsymbol{\theta}$. The function to be minimized is the *variational free energy* of the generic distribution Q:

[5] Mind, the Ising model is defined by a Hamiltonian:

$$H = \sum_{i<j} \sigma_i \sigma_j J_{ij} \tag{1.7}$$

being J a real, symmetric matrix, and $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_N)$, $\sigma_i = -1, 1$.

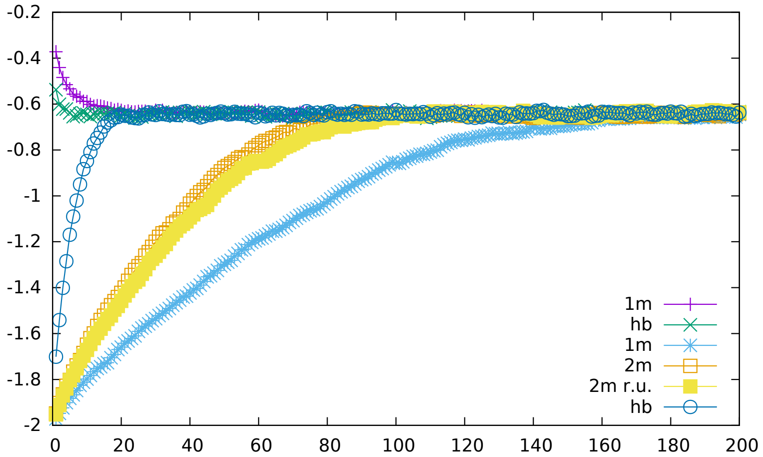

Figure 1.1: Energy vs. number of MC sweeps for the $q = 10$ 2D Potts model in the square lattice with periodic boundary conditions at $\beta = 1.24$, using several MC algorithms (Metropolis, heatbath, 2-hit Metropolis, 2-hit with random sweeps), and starting from ordered and disordered configurations (find the details of the simulation in the folder Potts/beta1.24 of [Ibanez-Berganza, 2016]).

$$\beta\tilde{F}[Q,\beta,J] = \beta\langle E[\mathbf{x},J]\rangle_Q - S[Q] = \tag{1.10}$$

$$= \langle \ln\frac{Q}{\mathcal{L}}\rangle_Q - \ln Z(\beta,J) = \tag{1.11}$$

$$= \beta F(\beta,J) + \mathrm{KL}[Q,\mathcal{L}] = \beta\tilde{F}[\mathcal{L},\beta,J] + \mathrm{KL}[Q,\mathcal{L}] \tag{1.12}$$

where $S[A] = -\langle \ln A\rangle_A$ is the entropy of the generic distribution A, $\mathrm{KL}(A,B) = \langle \ln(A/B)\rangle_A$ is the *relative entropy* (or Kullback-Leibler divergence) between the distributions A and B, and $F = -(1/\beta)\ln Z$ is the free energy (or the free energy functional of the target distribution, $\mathcal{L}$). According to Gibbs' inequality, the difference between variational and true free energies is $\Delta \geq 0$, the equality being satisfied only when the approximation turns exact. In other words, the minimization of $\tilde{F}$ provides an upper bound to F at the corresponding value of (β,J).

Thermodynamics of the Ising model by variational free energy minimisation. As an illustration of this variational scheme, we consider the Ising model as a particular case. The configuration space is $x_i \in \{-1,1\}$, and the energy

$$E[\mathbf{x},J] = -\frac{1}{2}\mathbf{x}^\dagger J\mathbf{x} - \mathbf{h}^\dagger\mathbf{x} \tag{1.13}$$

where, J is a real symmetric matrix with zero diagonal, $\mathbf{h}$ is a real vector (the explicit dependence of E in $\mathbf{h}$ will be ommitted in this section, absorbed in J).

If the distribution Q is factorizable in its variables, the calculation of the variational free energy (1.10) can be easily carried out (while the calculation of F requires a sum with 2^n terms). One supposes an exponential family $\boldsymbol{\theta} = \mathbf{a} = \{a_i\}_{i=1}^n$:

$$Q(\mathbf{x},\mathbf{a}) = \frac{e^{\sum_{i=1}^n x_i a_i}}{Z_Q}. \tag{1.14}$$

The probability of the i-th degree of freedom to be $+1$ is $q_i = 1/(1+e^{-2a_i})$. Being Q factorizable, its entropy is (check!) the sum of 1-particle entropies: $S[Q] = \sum_{i=1}^n h_2(q_i)$ where $h_2(y) = -y\ln y - (1-y)\ln(1-y)$. On the other hand, since the distribution Q is factorizable, the expectation value of the energy amounts to:

$$\langle E[\mathbf{x},J]\rangle_Q = -\sum_{i,j=1}^n \frac{1}{2}J_{ij}\bar{x}_i\bar{x}_j - \sum_{i=1}^n h_i\bar{x}_i \tag{1.15}$$

where $\bar{x}_i = 2q_i - 1 = \tanh(a_i)$ is the expectation value of x_i under Q.

Derivating the variational free energy, (1.10), with respect to a_i and equating to zero leads (check!) to the set of coupled equations:[6]

$$\begin{gathered} a_i = \beta\;[\sum_{j=1}^{n} J_{ij}\bar{x}_j + h_i] \\ \bar{x}_j = \tanh a_j \end{gathered} . \tag{1.16}$$

This is the mean field solution of the Ising model: a_i and $\bar{x}_i$ are the mean field and the magnetization of spin i, respectively. Given J and β, a solution (one of the in principle many possible solutions) can be obtained by assigning initial values to $\{\bar{x}_i\}_i$ and iterating the precedent equations, asynchronously (one particle at once) and indefinitely.

In the particular case of the Ising ferromagnent in a graph with coordination number $C = \sum_j J_{ij}$, these equations reduce to

$$a = \beta\;(CJ\bar{x} + h) \qquad \bar{x} = \tanh a \tag{1.17}$$

the solution of which is shown in Fig. 1.2 with $C = 4$, compared with the Onsager solution: $\bar{m} = (1 - \sinh(2\beta)^{-4})^{1/8}$ for $\beta > \beta' = \ln(1 + 2^{1/2})/2$, $\bar{m} = 0$ otherwise.

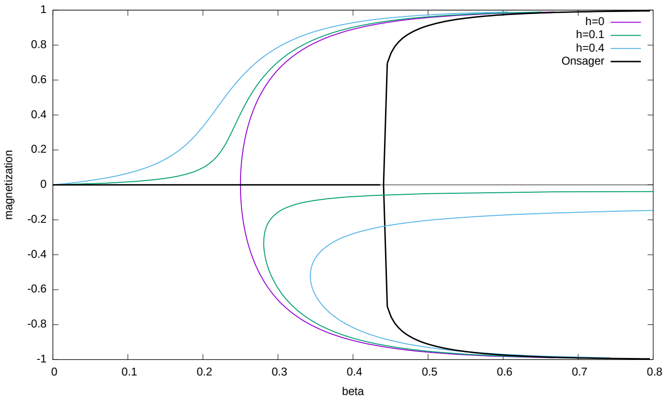

Figure 1.2: $\bar{m}(\beta)$.

***Exercise 12.** *Demonstrate Gibbs inequality, or* $\langle \ln(q/p) \rangle_q \geq 0$*, the equality being for* $q = p$ *only (it is enough to use the inequality* $\ln x \leq x - 1$*).*

***Exercise 13.** *Deduce the form of the variational free energy of the Ising model according to Q. Thus, obtain the variational solution of the Ising model in mean field approximation, eq. 1.16.*

Exercise 14. *Taylor-expand the function* $\bar{m}$ *in* β *around* $\beta > \beta'$ *and obtain the "*$\boldsymbol{\beta}$*" critical exponent* $\langle m \rangle \sim (\beta - \beta')^{\boldsymbol{\beta}}$ *corresponding to the Ising model in 2D (exact solution) and in mean field approximation (and obtain* 1/8 *and* 1/2*, respectively).*

[6] Mind that $\partial \cdot / \partial a_i = (\partial \bar{x}_i / \partial a_i) \partial \cdot / \partial \bar{x}_i$ and that $\partial \cdot / \partial a_i = (1/2)(\partial \bar{x}_i / \partial a_i) \partial \cdot / \partial q_i$ (being $\partial \bar{x}_m / \partial a_m = 1 - \bar{x}_m^2$).

2 Inference. The inverse problem in examples

2.1 Bayesian estimators

We remind Bayes equation:

$$P(\theta|D) = \frac{\mathcal{L}(D|\theta)\pi(\theta)}{\mathcal{E}(D)} \tag{2.1}$$

where θ are the hypothesis, D are the data, $P(\theta|D)$ is the posterior probability, $\mathcal{L}(D|\theta)$ is the data likelihood probability, $\pi(\theta)$ is the prior probability of hypothesis θ and $\mathcal{E}(D) = \sum_\theta \mathcal{L}(D|\theta)\pi(\theta)$ is the marginal likelihood or the evidence. Bayes equation follows from the definition of conditional probability: $p(A|B) = P(A, B)/P_1(B)$.

Given the data, a *Bayesian estimator* for the hypothesis, $\hat{\theta}$, is a value of the hypothesis minimizing the expectation $\langle R(\theta, \theta')\rangle_{P(\theta|D)}$ over the posterior of a given function R [called Bayes risk]. The Bayesian estimator corresponding to the mean square error as the Bayes risk is the average over the posterior: $\hat{\theta}(D) = \sum_\theta \theta P(\theta|D)$. An alternative estimator is the *Maximum A Posteriori* (MAP) estimator, or $\hat{\theta} = \arg\max_\theta P(\theta|D)$. In absence of any *a priori* information, when the prior probabilities are constants, the MAP estimator reduces to the *Maximum Likelihood* (ML) estimator $\hat{\theta} = \arg\max_\theta \mathcal{L}(D|\theta)$.

2.2 Maximum likelihood infering a Gaussian distribution

We consider n points $D = \{x_i\}_{i=1}^n$ identically, normally distributed. One can infer the mean and variance of the normal distribution by maximizing the log-likelihood with respect to them:

$$\ln \mathcal{L}(D|\mu, \sigma) = -n \ln[(2\pi)^{1/2}\sigma] - [n(\mu - \bar{x})^2 + S]/(2\sigma^2) \tag{2.2}$$

where $\bar{x}$ is the empirical average and $S = \sum_{i=1}^n (x_i - \bar{x})^2$. The likelihood can be described in terms of the functionals $S, \bar{x}$ of the data only, which recive the name of *sufficient statistics*. Differentiating the likelihood with respect to μ and σ leads to the ML estimators which jointly maximize the likelihood:

$$\mu^* = \bar{x} \tag{2.3}$$

$$\sigma^{*2} = n^{-1}S \tag{2.4}$$

Furthermore, the distribution of the likelihood of μ around its ML estimator μ^* is a normal distribution with standard deviation $\sigma\ n^{-1/2}$ (a particular instance of the central limit theorem) and the standard deviation of the likelihood distribution of $\ln \sigma$ is $(2n)^{-1/2}$.

While the resulting ML estimator for the mean is an unbiased estimator[7], the resulting ML estimator for σ results to be a biased estimator (check!). The unbiased estimator is obtained by *marginalizing* the likelihood with respect to the mean:

$$\mathcal{L}(D|\sigma) = \int_{-\infty}^{\infty} \mathrm{d}\mu\, \mathcal{L}(D|\mu, \sigma)\pi(\mu) \tag{2.5}$$

$$\ln \mathcal{L}(D|\sigma) = -n \ln((2\pi)^{1/2}\sigma) - S/(2\sigma^2) + \ln((2\pi/n)^{1/2}\sigma/\sigma_\mu) \tag{2.6}$$

the factor σ_μ^{-1} is the prior probability of μ (it results (check it!) as the leading approximation for a Gaussian prior for the average, with mean and variance μ_0 and σ_μ^2, in the limit of very large variance σ_μ^2). The ML estimator for σ^2, $\sigma^{*2} = \arg\max_{\sigma^2} \mathcal{L}(D|\sigma)$ results to be (check!):

$$\sigma^{*2} = S/(n-1) \tag{2.7}$$

***Exercise 15.** *Deduce equation (2.7).*

Exercise 16. *Obtain the marginal probability distribution for the average, $\mathcal{L}(D|\mu)$. What is such distribution?*

[7] an unbiased estimator E of a quantity Q being such that $\langle E[D]\rangle_D = \langle Q\rangle$, where $\langle\cdot\rangle$ is the average over the true distribution and $\langle\cdot\rangle_D$ is the average over many realizations of the data D, generated according to the true distribution.

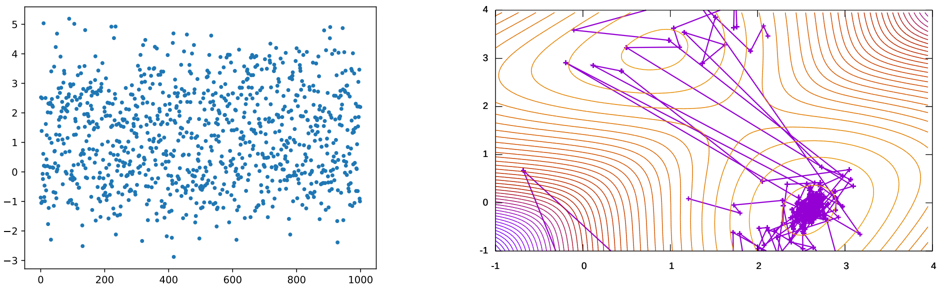

Figure 2.1: Left: $n = 10^3$ data extracted from the probability distribution (2.8), with $K = 2$, $p_1 = 1/2$, $\mu_1 = 2.5$, $\mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$. Right: successive values of the parameters $\theta_{1,2} = \mu_{1,2}$ sampled from a MC Metropolis algorithm, in the (μ_1, μ_2) space (the likelihood is represented by contour iso-likelihood lines) (see the details, the algorithm scripts and the data in /BayesianMixture/Metropolis/, in [Ibanez-Berganza, 2016]). The probabilities p_1, $1 - p_1$ in the MC algorithm are also inferred. Note that only the absolute maximum corresponds to (but do not coincide with) the true parameters used to generate the data.

2.3 Inferring a mixture of Gaussian distributions

Mixtures of probability distributions. Consider n data $\mathbf{x} = \{x_i\}_{i=1}^n$ generated with a mixture of K probability distributions, each data generated from the j-th distribution, f_j, with parameters θ_j with probability p_j, being $\sum_{j=1}^K p_j = 1$, $\mathbf{p} = \{p_j\}_{j=1}^K$, $\boldsymbol{\theta} = \{\theta_j\}_{j=1}^K$. The likelihood can be written as (from now on we will absorb $\mathbf{p}$ into θ):

$$\mathcal{L}(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{j=1}^{K} p_j f(x_i|\theta_j) \right] . \tag{2.8}$$

Although the likelihood (2.8) can be evaluated in $\mathcal{O}[Kn^2]$, there are K^n terms in the sum, so that the direct evaluation of Bayesian estimators is not feasible.

Monte Carlo estimation of the likelihood estimator One is interested in an estimator for the parameters $\mathbf{p}$ and $\boldsymbol{\theta}$, i.e., one looks for $\langle\boldsymbol{\theta}\rangle_{\mathcal{L}(\mathbf{x}|\boldsymbol{\theta})}$. A possibility is to implement a Monte Carlo chain whose stationary distribution in $\boldsymbol{\theta}$ is $\mathcal{L}(\mathbf{x}|\boldsymbol{\theta})$.

One possibility is to implement a Monte Carlo algorithm of the so called Metropolis type (see the Appendix A). One chooses an algorithm (see the Appendix A) based on a probability transition matrix between two states, $P[\boldsymbol{\theta} \to \boldsymbol{\theta}']$ equal to $\min\{1, \mathcal{L}(\mathbf{x}|\boldsymbol{\theta}')/\mathcal{L}(\mathbf{x}|\boldsymbol{\theta})\}$. It can be proved (see references in [Ibanez-Berganza, 2016]) that such transition rule satisfies detailed balance and it is aperiodic; it follows that the stationary probability distribution on the $\boldsymbol{\theta}$'s is given by $\mathcal{L}(\mathbf{x}|\boldsymbol{\theta})$ and the Bayesian estimator (the average $\langle\boldsymbol{\theta}\rangle_{\mathcal{L}(\mathbf{x}|\boldsymbol{\theta})}$) is given by the (long-time) average of the sequence of resulting $\boldsymbol{\theta}$'s via eq. (1.5).

Numerical example Let us show a numerical example. We will consider the simple case of a mixture of Gaussians: $\mathcal{N}(\theta_i, 1)$, i.e., $f(x_i|\theta_j) = (2\pi)^{-1/2}\exp(-(x_i - \theta_j)^2/2)$ with fixed variance. In particular, the simple $K = 2$ case such that the only hypothesis to be inferred from the data $\mathbf{x}$ are the average of the distribution: $\boldsymbol{\theta} = (\mu_1, \mu_2)$. The prior probabilities are supposed to be fixed and known, $p_j = 1/2$, and so the standard deviations, $\sigma_1 = \sigma_2 = 1$. The probability distributions are assumed to be Gaussian, $f = \mathcal{N}$.

Fig. 2.1 shows one hundred points $n = 100$ generated with $\mathcal{L}(\mathbf{x}|\boldsymbol{\theta})$ (right), with known (to be inferred a posteriori) averages $\mu_1 = 2.5$, $\mu_2 = 0$. The left figure shows the Likelihood function landscape in the 2-dimensional $\boldsymbol{\theta} = (\mu_1, \mu_2)$ space, along with a series of $\boldsymbol{\theta}$ states generated with a Metropolis algorithm.

2.4 Maximum likelihood inferring the coupling parameters of a statistical model.

Consider the inference of the parameters J of a Boltzmann distribution $P(\mathbf{x}|J)$ corresponding to a pairwise Hamiltonian whose (real and symmetric) interaction matrix is J:

$$P(\boldsymbol{\sigma}|J) = \frac{1}{Z} \exp\left[-\frac{1}{2}\sum_{i,j} \sigma_i \sigma_j J_{ij}\right] \tag{2.9}$$

from a finite set of configurations $\{\boldsymbol{\sigma}_s\}_{s=1}^{S}$ sampled from P. The ML inference consists in the maximisation of *the likelihood of the joint set of configurations*, $\prod_s P(\boldsymbol{\sigma}^{(s)}|J)$. Differentiating this quantity with respect to the model parameters and equating them to zero results in:

$$\frac{\partial}{\partial J_{ij}} \sum_s \ln P(\boldsymbol{\sigma}^{(s)}) = 0 \implies \langle \sigma_i \sigma_j \rangle_{\mathrm{e}} = \langle \sigma_i \sigma_j \rangle_P \tag{2.10}$$

How to search in the J's space a solution to the above equation? One could implement a recursive dynamics for the J matrix in a discrete time t so that the updating $J(t+1) = J(t) + \delta(t)$, where $\delta(t)$ is the matrix:

$$\delta_{ij}(t) = \eta \left.\frac{\partial}{\partial J_{ij}}\right|_{J(t)} \sum_k \ln P(\boldsymbol{\sigma}^{(k)}) = \eta \left[-\langle \sigma_i \sigma_j \rangle_{\mathrm{e}} + \langle \sigma_i \sigma_j \rangle_{P(t)}\right] \tag{2.11}$$

we correct, hence, the J's with a quantity proportional to the violation of the equality (2.10). The inverse problem is at least as difficult as the direct problem. In the general case, one has to search in the space of the parameters J, but each time one evaluates $\langle O_k \rangle_{P(t)}$ one has to solve a direct problem for the current value of the parameters $J(t)$.

In the following, we present a particularly simple situation in which the inverse problem can be solved easily and analytically, and an approximation overcoming the aforementioned recursive search in the parameters space in the case of Ising variables.

Exercise 17. *In an interacting model, as the one defined by (2.9) (and such that the coupling strengths (more precisely, βJ) are not very small), does the correlation $\langle x_i x_j \rangle$ necessarily vanish if $J_{ij} = 0$? The question can be considered in the light of a perturbative series of e^{-H}, a Mayer expansion): demonstrate that, at first order in H (say, in β), two Ising spins are correlated only if they interact. Convince yourself that, at higher orders in H, this is not necessarily true.*

Exercise 18. *Does a model with pairwise interactions, as the one defined by (2.9), present necessarily null higher-tan-2-point (non-connected) correlations $\langle x_{i_1} \cdots x_{i_{p>2}} \rangle$? Does a model with pairwise interactions present necessarily null higher-tan-2-point cumulants (or connected correlations) $\langle\langle x_{i_1} \cdots x_{i_{p>2}} \rangle\rangle$ (see appendix B)? Under what conditions a statistical model defined by a $p = 2$ (pairwise) interacting Hamiltonian presents vanishing (connected) correlations of all orders $q > 2$? (see appendix B for a definition of connected correlation).*

2.4.1 Inference of a model with real degrees of freedom (from pairwise correlations)

We will solve the direct problem of a d-dimensional model with real (positive and negitve) degrees of freedom, $\mathbf{x} = (x_i)_{i=1}^{d}$, with a Boltzmann probability density in the canonical ensemble at temperature $= 1$ and a Hamiltonian given by the quadratic form J (a symmetric, positive-definite matrix):

$$P(\mathbf{x}) = \frac{1}{Z} \exp\left[-\frac{1}{2}\mathbf{x}^\dagger J \mathbf{x} - \mathbf{h}\cdot\mathbf{x}\right] \tag{2.12}$$

$Z = (2\pi)^{d/2} \det(J)^{-1/2}$ is the partition function, normalizing P. The connected two-point correlator according to P, $\langle\langle x_i x_j \rangle\rangle_P \equiv \langle x_i x_j \rangle_P - \langle x_i \rangle_P \langle x_j \rangle_P$, and the averages $\langle x_i \rangle_P$ according to P are:

$$\begin{aligned} \langle\langle x_i x_j \rangle\rangle_P = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle &= (J^{-1})_{ij} \\ \langle x_i \rangle_P &= (J^{-1}\mathbf{h})_i \end{aligned} \tag{2.13}$$

So that the inverse problem is analytically solvable. If C is the (connected) experimental correlation matrix, then the inverse problem is simply $J = C^{-1}$.

***Exercise 19.** *Demonstrate the cumulant expansion (see appendix B) in the $p = 2$ case:*

$$\langle x_i x_j\rangle - \langle x_i\rangle\langle x_j\rangle = \left.\frac{\partial^2}{\partial_{u_i}\partial_{u_j}}\right|_{\mathbf{u}=\mathbf{0}} \ln Z(\mathbf{u}) \tag{2.14}$$

where $\ln Z(\mathbf{u})$ is the generating function:

$$\ln Z(\mathbf{u}) = \int \mathrm{d}\mathbf{x} \exp(-H(\mathbf{x}) + \mathbf{x}\cdot\mathbf{u}). \tag{2.15}$$

***Exercise 20.** *Demonstrate the equations 2.13, and the form of the partition function of the Gaussian model $Z = (2\pi)^{d/2}\det(J)^{-1/2}$ in the case of zero averages, and $Z = (2\pi)^{d/2}\det(J)^{-1/2}\exp(\frac{1}{2}\mathbf{h}^\dagger J^{-1}\mathbf{h})$ for non-vanishing $\mathbf{h}$'s (see appendix C).*

2.4.2 Inference of Ising degrees of freedom in the linear response approximation (from pairwise correlations)

We will apply *linear response theory* to the maximum entropy problem, particularized for the Ising model. The mean field solution of sec. 1.3 is such that the average 2-point correlator vanishes. However, one can consider the general expression (check it!, and see appendix B for a generalisation):

$$\langle x_i\rangle = -\frac{\mathrm{d}F}{\mathrm{d}h_i} \tag{2.16}$$

$$\langle x_i x_j\rangle = -\frac{1}{\beta}\frac{\mathrm{d}^2 F}{\mathrm{d}h_i \mathrm{d}h_j} + \langle x_i\rangle\langle x_j\rangle \tag{2.17}$$

where $F = -\ln Z/\beta$ is the free energy. One can now approximate F by the minimum of $\tilde{F}$ in sec. 3, that will be called $\tilde{F}(\beta, J)$ (in other words, $\tilde{F}(\beta, J)$ is what before was $\tilde{F}(Q, \beta, J)$ with $Q(\mathbf{x}, \mathbf{a})$ evaluated in the a_i and $\bar{x}_i$ satisfying the equations 1.16). This approximation, $F \simeq \tilde{F}$, will be called *linear response approximation* [Kappen and Rodríguez, 1998]. One can see that (check!):

$$\frac{\mathrm{d}\tilde{F}}{\mathrm{d}h_i} = \frac{\partial\tilde{F}}{\partial h_i} + \sum_{j=1}^{n}\frac{\mathrm{d}a_j}{\mathrm{d}h_i}\frac{\partial\tilde{F}}{\partial a_j} \tag{2.18}$$

$$\langle x_i\rangle \simeq \bar{x}_i \tag{2.19}$$

Note that the second term of the first equation vanishes, since $\tilde{F}$ has been chosen as the minimum w.r.t. the a's. In linear response approximation, the averages are as in the bare mean field approximation of sec. 3. Oppositely, the correlations:

$$\begin{aligned}\langle x_i x_j\rangle - \langle x_i\rangle\langle x_j\rangle = \frac{1}{\beta}A_{ij}, \qquad A_{ij} \equiv \frac{\mathrm{d}\bar{x}_j}{\mathrm{d}h_i}\\ (A^{-1})_{ij} = \delta_{i,j}\frac{1}{\beta(1-\bar{x}_i^2)} - J_{ij}\end{aligned} \tag{2.20}$$

The first line is a consequence of eq. (2.19), while the second line can be demonstrated (check!) derivating w.r.t h_i the equation for $\bar{x}_j = \tanh(a_i)$, 1.16 (hint: use $\mathrm{d}\bar{x}_i/\mathrm{d}h_j = (\mathrm{d}\tanh(a_i)/\mathrm{d}a_i)\,(\mathrm{d}a_i/\mathrm{d}h_j)$). Hence, in linear response approximation, the connected two-point correlation matrix is (except by the diagonal) minus the inverse of the interaction matrix.

Exercise 21. *Demonstrate the linear response equations, 2.20.*

***Exercise 22.** *Why $J = C^{-1}$ in equation (2.13), but $J = -C^{-1}$ in (2.20)? Is it a question of sign convention?*

Exercise 23. *But, more importantly, how is it possible that the direct problem turns equal, $C = \pm J^{-1}$, in both models (see (2.20,2.13)), that are so different (the Ising model degree of freedom is $x_i \in \{-1, 1\}$, while in the Gaussian model it is $x_i \in \mathbb{R}$)?*

Effective field theory of the Ising model. In the continuum limit (vanishing lattice space), the Ising model with Hamiltonian $H = -\sum_{i<j} x_i x_j J_{ij}$, $x_i = -1, 1$, $i = 1, \ldots, N$, can be effectively described by a theory in terms of scalar continuum fields ϕ_i, $i = 1, \ldots, N$. The theory is defined by the partition function (see, for example, [Amit and Martin-Mayor, 2005, Mussardo, 2010]):

$$Z = \int [\mathrm{d}\phi] \exp\left(-\frac{1}{2} \boldsymbol{\phi}^\dagger \cdot J \cdot \boldsymbol{\phi} + \sum_i \ln[\cosh((2J \cdot \boldsymbol{\phi})_i)] \right) \tag{2.21}$$

Keeping the bilinear form in the exponent is equivalent to the mean-field approximation.

Exercise 24. *Consider the inverse problem, in pairwise approximation, of a model with real degrees of freedom. Demonstrate that the expression for the data log-likelihood is* $\langle \ln P(\cdot|J^*) \rangle_{\mathrm{e}} = -(n/2 + 1) \ln(2\pi) + (1/2) \sum_{i=1}^{n} \ln \epsilon_i$*, where* ϵ_i *are the eigenvalues of matrix* J *(see, if you want, appendix C). Hence, a large value of the data likelihood is equivalent to an "accurate fit" (the variances along the principal axes induced by the model,* ϵ_i^{-1}*, are small). In other words, a lower generalised entropy (averaged over the data) means that there is few ambiguity in the description of the data by* P*.*

3 Maximum entropy inference

3.1 General formulation

Consider an n-body *configuration* (or *phase*) *space* $\Sigma = \Sigma_1^{\otimes n}$, whose configurations are called $\mathbf{x} = (x_i)_{i=1}^n \in \Sigma$. Suppose that one has M experimental measurements of a set of K observables (whose corresponding operator in Σ is called $O_k : \Sigma \to \mathbb{R}$). The experimental averages are called $\langle O_k \rangle_{\rm e} = \frac{1}{M}\sum_{i=1}^M O_k^{(i)}$, where $O_k^{(i)} = O(\mathbf{x}^{(i)})$ is the i-th experimental result (a number), $i = 1, \cdots, M$, of the k-th observable applied to the sample $\mathbf{x}^{(i)}$.

The *maximum entropy* approach provides the *most probable model*, or probability distribution (or likelihood), $P(\mathbf{x}|\boldsymbol{\lambda})$, $\mathbf{x} \in \Sigma$, which is consistent with the experimental observations $\langle O_k \rangle_{\rm e}$ (called *sufficient statistics*), in the sense that it is constrained to reproduce them by construction:

$$\langle O_k \rangle_P = \langle O_k \rangle_{\rm e} \tag{3.1}$$

(where $\langle \cdot \rangle = {\rm tr}_{\mathbf{x}} \cdot P(\mathbf{x}|\boldsymbol{\lambda})$).

In other words the *maximum entropy* distribution $P_{\rm me}$ is the most random, or less structured distribution subject to the constraint (3.1), and to no other constraint. The probability distribution with maximum entropy P results from the extremum condition of the so called generalized entropy:

$$\mathcal{S}[P] \equiv S[P] + \sum_{k=1}^K \lambda_k (\langle O_k \rangle_P - \langle O_k \rangle_{\rm e}). \tag{3.2}$$

The maximum of the generalised entropy is the maximum of the entropy of the distribution, when it is subject to the constraints (3.1). Functional-derivating (3.2) with respect to $P(\mathbf{x})$ and equating to zero (the normalization may be ensured by a further Lagrange multiplier, λ_0) results in (check!):

$$P_{\rm me}(\mathbf{x}) = \frac{1}{Z(\boldsymbol{\lambda})} \exp\left[\sum_{k=1}^K \lambda_k O_k(\mathbf{x})\right] \tag{3.3}$$

$Z(\boldsymbol{\lambda}) = \exp(\lambda_0 - 1)$ being the normalizing constant. The maximum entropy probability distribution is, hence, a Boltzmann distribution in the canonical ensemble at temperature $= 1$, with effective Hamiltonian $\mathcal{H} = -\sum_k \lambda_k O_k$. It is important to remark that no assumption at all has been done about thermal equilibrium, ergodicity, nor about the existence of an effective interaction in energy units: the Boltzmann form is a consequence of the maximum entropy assumption –reflecting, rather, *absence* of hypothesis– of a probability distribution subject to constraints. In [Jaynes, 1957a]'s words:

> If in addition we reinterpret the prediction problem of statistical mechanics in the subjective sense, we can derive the usual relations in a very elementary way without any consideration of ensembles or appeal to the usual arguments concerning ergodicity or equal a priori probabilities. The principles and mathematical methods of statistical mechanics are seen to be of much more general applicability than conventional arguments would lead one to suppose. In the problem of prediction, the maximization of entropy is not an application of a law of physics, but merely a method of reasoning which ensures that no unconscious arbitrary assumptions have been introduced.

What about the values of the Lagrange multipliers λ's in (3.3)? They are determined by imposing the constraints that $P_{\rm me}$ is required to satisfy, (3.1). This is equivalent (check!) to the optimisation of the function (3.2) with respect to the λ's (its correct value can be shown to be a *minimum*).

Relationship with maximum likelihood. Notice that, alternatively, the minimization of the generalized entropy is equivalent to the maximization of the experimental average of the likelihood (from eq. 3.2, check!):

$$\mathcal{S}[P] = \ln Z(\boldsymbol{\lambda}) - \sum_{k=1}^{K} \lambda_k \langle O_k \rangle_{\rm e}$$
$$= -\langle \ln P_{\rm me} \rangle_{\rm e} = (1/M) \sum_{m=1}^{M} \ln P(\mathbf{x}^{(m)}) \tag{3.4}$$

where $\mathbf{x}^{(m)}$ is the m-th experimental configuration.

In other words, the λ's are chosen by imposing (3.1) or, equivalently, by *minimising* (3.4, i.e., by maximising the *global, experimental likelihood* according to the model, P).

Thus, the parameters λ (called effective couplings) are obtained by maximum likelihood (minimum generalised entropy), once one has assumed (by maximum generalised entropy) that the most probable distribution has the form (3.3).[8]

Exercise 25. *Check that the database maximum likelihood recipe to fix the parameters is equivalent to the minimisation of the relative entropy among the database empirical histogram, $h(\mathbf{x}) = \sum_m \delta_{\mathbf{x}^{(m)},\mathbf{x}}$, and the inferring probability distribution function.*

***Exercise 26.** *Demonstrate eq. 3.3. Check that imposing the constraints (3.1) is equivalent to optimise the generalised entropy with respect to λ. Check that it is actually equivalent to minimise the generalised entropy (or to maximise the data likelihood).*

***Exercise 27.** *Maximum entropy? Minimum entropy? You said that the generalised entropy is to be maximised, to get to 3.3. Afterwards, that it is minimised (see equation 3.4, and exercise 26). Where is the catch?*

3.2 Maximum entropy inference with pairwise correlations

Suppose one wants to perform maximum entropy inference in a system with general degrees of freedom $x_i \in \Sigma_1$, given that the observables O of sec. 3.1 are averages and correlators (i. e., 1- and 2-point operators respectively): x_i, $x_i x_j$. The maximum entropy probability distribution on Σ results to be (c. f. (3.3)):

$$P_{\rm me}(\mathbf{x}|J,\mathbf{h}) = \frac{1}{Z(J,\mathbf{h})} \exp\left[\sum_{i,j=1}^{n} J_{ij} x_i x_j + \sum_{i=1}^{n} h_i x_i\right] \tag{3.5}$$

i. e., a Boltzmann distribution at inverse temperature $= 1$, with the couplings and fields J, $\mathbf{h}$ such that:

$$\langle x_i x_j \rangle_P = \langle x_i x_j \rangle_{\rm e} \qquad \langle x_i \rangle_P = \langle x_i \rangle_{\rm e} \tag{3.6}$$

where $\langle\cdot\rangle_{\rm e}$ refers to the experimental average. The problem is, in general, hard, when the evaluation of $\langle\cdot\rangle_P$, the direct problem, is not immediate.

Information provided by ME. Once one has solved the inverse problem (with the limitations given by the finiteness of the experimental data), one has access to information that was, in principle, not directly accessible from the data: the microscopic information provided by the effective interactions (typically different from the correlations, see before); the possibility of *recognizing or classifying* novel configurations or creating new ones (see before); the possibility of estimating new observables, $\langle O \rangle_P$ if they do not have been measured.

Sufficiency of the sufficient statistics. Indeed, a self-consistency test of the sufficiency of the sufficient statistics O_k is that of calculating different nontrivial observables according to P (different from the sufficient statistics, i.e., observables that P is not required to reproduce by construction), and comparing them with their experimental counterparts. Two-degree of freedom

[8] This can also be viewed (check!) as a minimization of the relative entropy (c.f. section 1.3), $KL[h,P]$, where h is the experimental histogram of the data, $h(\mathbf{x}) = \sum_m \delta_{\mathbf{x}^{(m)},\mathbf{x}}$. This is the maximum entropy-maximum likelihood-minimum free energy functional relationship.

experimental correlations agree with those according to P by construction. One could perform the test with fourth-order correlations: as far as $\langle (x_i x_j)(x_m x_n)\rangle_P \simeq \langle (x_i x_j)(x_m x_n)\rangle_{\rm e}$, the pairwise maximum entropy approximation is correct. In general, a recipe is to use p-th order correlations as sufficient statistics, with p such that the $p+1$-th order experimental correlations are reproduced by $P_{\rm me}$.

Errors in ME. There are actually (at least) three sources of errors in maximum entropy inference. (1) The choice of the operators O and their number could lead to a unfaithful generative model of the data. Moreover, choosing to many operators in the sufficient statistics could lead to *overfitting* (the excessive dependence of $P_{\rm me}$ on the sample non-significant details). (2) Even in the ideal case that these are correct (that the functional form of the generative model from which the data has been extracted coincides with that of $P_{\rm me}$), one can infer poorly simply due to the ambiguity induced by the finiteness of the input data. (3) Even in the mentioned ideal case, the data likelihood maximisation in the λ parameters may be a difficult problem, due to the presence of many local maxima of the likelihood in the λ space [Nguyen et al., 2017].

3.2.1 Applications of maximum entropy inference

We mention different applications pairwise maximum entropy inference.

1. **Inferring neural activity.** In recent years, binary pairwise models have been extensively used as parametric models for studying the statistics of spike trains of neuronal populations and for inferring neuronal functional connectivities [Schneidman et al., 2006, Shlens et al., 2006, Tang et al., 2008, Shlens et al., 2006].

   »Using maximum entropy methods from statistical mechanics, we show that pairwise and adjacent interactions accurately accounted for the structure and prevalence of multi-neuron firing patterns, explaining $\sim 98\%$ of the departures from statistical independence in parasol cells and $\sim 99\%$ of the departures that were reproducible in repeated measurements. [Shlens et al., 2006].

   »Here we show, in the vertebrate retina, that weak correlations between pairs of neurons coexist with strongly collective behaviour in the responses of ten or more neurons. We find that this collective behaviour is described quantitatively by models that capture the observed pairwise correlations but assume no higher-order interactions. These maximum entropy models are equivalent to Ising models, and predict that larger networks are completely dominated by correlation effects. [Schneidman et al., 2006]

   See much more on the mean field approximation applied to Ising pairwise inferring in [Roudi et al., 2009, Nguyen et al., 2017, Berg, 2017].

2. In [Morcos et al., 2011] have used maximum entropy ot infer protein conformations and structures from the correlation between different amino-acid compositions at different sequence positions.

3. Inferring the effective interaction properties in flocks of birds.

   In reference [Cavagna et al., 2015] (see also [Bialek et al., 2012]) correlation between velocities of birds in a swarm are considered, but the correlations $\langle \mathbf{p}_i \mathbf{p}_j\rangle_{\rm e}$ are not taken among the i-th and j-th individuals, as this would not allow to measure several instancies of the correlations (since the birds move in time). Instead, as operators f, in the terminology of sec. 3, it is used $C(\{\mathbf{v}\}, d)$, the velocity correlation between between two birds at different topological distancies, $d = 1, 2, \cdots, n$:

$$C(\{\mathbf{v}\}, d) = \frac{1}{n}\sum_{i,j=1}^{n} \mathbf{v}_i \cdot \mathbf{v}_j \delta_{D_{ij},d} \tag{3.7}$$

   where D_{ij} is a non-symmetric matrix defined such that $D_{ij} = m$ if j is the m-th nearest neighbor of i in 3D space. The effective energy of the maximum entropy distribution P is (check!):

$$-\sum_{d=1}^{N} J_d \, C(\{v\}, d) = -\sum_{i,j} J_{D_{ij}} \, \mathbf{v}_i \cdot \mathbf{v}_j \tag{3.8}$$

The analytical expression of the partition function:

$$\ln Z(J) = -\sum_{j=2}^{n} \ln a_j + n \sum_{d} J(d) \tag{3.9}$$

makes possible the maximization of the log likelihood:

$$\ln P = \langle \ln Z(J) + n \sum_{d} J(d) C(\{s\}, d) \, \rangle_{\mathrm{e}} \tag{3.10}$$

with respect to the function J. In this equation, we have stressed that the partition function has to be averaged w.r.t. the experimental sample, since the graph J_{ij} dependence (which varies from sample to sample of the swarm) of Z, c.f. (3.9).

4. In [Bethge and Berens, 2008], maximum entropy has been applied to the statistical study of pixel intensities in natural images.

5. In [Sakellariou et al., 2016], an improvement of mean field maximum entropy, called pseudo-likelihood inference, is used to capture statistics of melodies in music.

6. Consider a set of S human subjects selecting the image of a face among a set of human faces. The faces are codified in a D-dimensional vector $\mathbf{x}$, describing some set of facial distances (determined by the mutual distances between the spatial coordinates some landmarks, or points that can be unambiguosly determined on each face). We have, hence $\mathbf{x}^{(s)}$, $s = 1, \ldots, S$ D-dimensional experimental configurations. If the $\mathbf{x}$'s determine, more precisely, the fluctuations of the aforementioned distances with respect to their average (so that $\langle \mathbf{x} \rangle_{\mathrm{e}} = \mathbf{0}$, the $\{\mathbf{x}^{(s)}\}_{s=1}^{S}$ are real (positive and negative) D-dimensional vectors with zero mean on every component. One is, in principle, legitimated to apply the maximum entropy method using pairwise correlations as sufficient statistics. The validity of this approach may be *a posteriori* assessed. The dataset is in this way described by a probability distribution in the space of facial coordinates $P(\mathbf{x})$, such that the experimental and theoretical two-facial distance correlations coincide: $\langle x_i x_j \rangle_P = \langle x_i x_j \rangle_{\mathrm{e}}$. In [Ibáñez-Berganza et al., 2019], it is shown how the resulting ME probability distribution: (1) provides a faithful description of the dataset, as shown by its ability as a *classifier*; (2) reproduces observables (different from the sufficient statistics $\langle x_i x_j \rangle_{\mathrm{e}}$); (3) provides novel information regarding the human cognitive process of facial preference and discrimination.

4 Elements of Bayesian model selection

In a Bayesian statistics framework, one has a tool to discriminate among different models of inference with different *complexity*, i.e., number of parameters [MacKay, 2003]. Such a discrimination is called *model selection.*

In general, given the experimental data, $\mathcal{D}$, and the model $\mathcal{M}$ (say, the functional form of the likelihood $\mathcal{L}(\mathcal{D}|J,\mathcal{M})$, where J are the model parameters, depending on $\mathcal{M}$). The higher the likelihood, the better the description of the data given the model. When various models $\mathcal{M}$ are defined, a first form of model selection is to evaluate the maximum likelihood of the data, $\mathcal{L}(\mathcal{D}|J^*,\mathcal{M})$ for different models, being J^* the Maximum Likelihood (ML) estimations given the data and each model $\mathcal{M}$. However, the ML estimation does not take into account the possible overfitting, nor the ability of the model to *generalise*, to learn relevant variations of the data. A mechanism of model selection taking into account the overfitting should evaluate as well the robustness of the inference results (J^*) with respect to fluctuations in the data (or, equivalently, in the inference parameters themselves, given the data).

In Bayesian statistical terms, to perform model selection one evaluates the *probability of a model, given the data*, something different from the ML probability (*the probability of the data given the most probable model*). The probability of a model given the data, a quantity called *data evidence*, accounts not only for the maximum likelihood of the data given a model, but also for the possible data overfitting, penalising the fine tuning of many model parameters. Consider the Bayes equation:

$$\mathcal{P}(J,\mathcal{M}|\mathcal{D})\mathcal{E}(\mathcal{D},\mathcal{M}) = \mathcal{L}(\mathcal{D}|J,\mathcal{M})\Pi(J,\mathcal{M}) \tag{4.1}$$

where $\mathcal{P}$, $\mathcal{E}$, $\mathcal{L}$, Π are, respectively, the posterior, the evidence, the likelihood and the prior probabilities, and $\mathcal{D}$ represents the data, $\mathcal{M}$ the model and J the model parameters (whose number depends on $\mathcal{M}$, since various models may exhibit different complexity). One is interested in evaluating the evidence of different models, to assess which is the most probable model given the data. In a Bayesian scheme, the model with higher evidence is the most probable model. The evidence takes the form (notice that the posterior probability is normalized in J) (we ommit the $\mathcal{M}$ index):

$$\mathcal{E}(\mathcal{D}) = \int [\mathrm{d}\, J]\mathcal{L}(\mathcal{D}|J)\Pi(J) \simeq \mathcal{L}(\mathcal{D}|J^*)\int [\mathrm{d}J] e^{-\frac{1}{2}\vec{\delta J}^\dagger\cdot nF(J^*)\cdot\vec{\delta J}}\Pi(J) \tag{4.2}$$

where $J^* = \arg\max_J \mathcal{L}(D|J)$ and we have supposed that $\mathcal{L}$ is Gaussianly distributed in J around its maximum J^* for small variations of around J^*. $\vec{\delta J}$ are a vector form (the multiple indices of J have been aligned) of $\delta J \equiv J - J^*$, n is the number of observations, $\mathcal{D} = \{\mathbf{x}\}_1^n$ and $F(J)$ is the Fisher matrix of $\mathcal{L}$ evaluated in J and averaged over the data, i.e.:

$$F(J) = -\langle \mathrm{Hess}\,\{\ln\mathcal{L}\}|_J\rangle_{\mathrm{e}} \tag{4.3}$$

or:

$$F(J)_{\mu\nu} = -\left\langle \left.\frac{\partial^2}{\partial J'_\mu \partial J'_\nu}\right|_J \ln\mathcal{L}(\mathbf{x}|J')\right\rangle_{\mathrm{e}} \tag{4.4}$$

where $\langle O\rangle_{\mathrm{e}} = (1/n)\sum_{i=1}^n O(\mathbf{x}^{(i)})$ and $\mathcal{L}(\mathcal{D}|J) = \prod_i \mathcal{L}(\mathbf{x}^{(i)}|J)$, in such a way that:

$$\ln\mathcal{L}(\mathcal{D},J) = \ln\mathcal{L}(\mathcal{D}|J^*) - \frac{1}{2}\vec{\delta J}^\dagger\cdot nF(J^*)\cdot\vec{\delta J} + \mathcal{O}[(\delta J)^4] \tag{4.5}$$

The evidence equation 4.2 takes into account not only the contribution of the maximum likelihood, but also the contribution to the likelihood for *all* the possible values of J weighted from their probability distribution, $\Pi(J)$. The integral in 4.2 is called the *Occam term*: models are advantaged if the likelihood keeps hight for most values of J in their space with measure $\mathrm{d}\mu_J = \mathrm{d}J\,\Pi(J)$, while they are penalized if they have to be fine-tuned to achieve a good result: if the function $\mathcal{L}(D|J)$ is low for values of $J \neq J^*$ in the J's space with measure $\mathrm{d}\mu_J$.

Supposing that the prior Π is locally linear in J around J^*, it is, by Gaussian integration:

$$\ln\mathcal{E}(\mathcal{D}) = \ln\mathcal{L}(\mathcal{D}|J^*) + \ln\Pi(J^*) + \frac{p}{2}\ln\frac{2\pi}{n} - \frac{1}{2}\ln\det F(J^*) \tag{4.6}$$

where p is the number of parameters, J_μ with $\mu = 1,\dots,p$.

Supposing, for simplicity, a uniform prior in a p-dimensional hyper-cube of length σ_J in every coordinate J_μ, it is $\Pi(J) = \sigma_J^{-p}$. Given the data, the standard deviation of $\mathcal{L}(\mathcal{D}|J)$ around J^* in the J space is $\det\{nF(J^*)\}^{-1/2}$. Supposing that $\sigma_{J|\mathcal{D}}$ is the standard deviation per coordinate J_μ, and that they are all of the same order, one can write $\det\{nF(J^*)\}^{-1/2} = \sigma_{J|\mathcal{D}}^p$, and:

$$\ln \mathcal{E}(\mathcal{D}) = \ln \mathcal{L}(\mathcal{D}|J^*) + \frac{p}{2}\ln 2\pi + p \ln \frac{\sigma_{J|\mathcal{D}}}{\sigma_J} \tag{4.7}$$

i.,e., the evidence increases when the likelihood increases (through the ML term), but it decreases (through the Occam term) when the inference parameters are too fine tuned $\sigma_{J|\mathcal{D}} \ll \sigma_J$ around its most probable value, and this effect is exponentially large in the number of parameters. Tipically, when the number of parameters increases, the likelihood, on the one hand, increases since the data is better fitted with more parameters. On the other hand, the Occam factor decreases (increases in absolute value) since, first of all, it is proportional to p and, second, inferring a larger number of parameters results in the parameters being more fine tuned around their most probable value, i.e., $\sigma_{J|\mathcal{D}}$ decreases.

To see better the dependence of $\mathcal{E}$ on p and n we can neglect all the terms of order $\mathcal{O}[1]$ in n in the saddle point approximation for $\mathcal{E}$, equation 4.6. It is:

$$\ln \mathcal{E}(\mathcal{D}) = n\langle \ln \mathcal{L}(\cdot|J^*)\rangle_{\mathrm{e}} - \frac{p}{2}\ln n + \mathcal{O}[n^0] \tag{4.8}$$

For large number of observations, n, maximising the data evidence (in its saddle-point approximation, equation 4.6) is equivalent to minimising the so called Bayesian Information Criterion (BIC):

$$B = p \ln n - 2 \ln \mathcal{L}(\mathcal{D}|J^*, \mathcal{M}) \tag{4.9}$$

where, we recall, J^* is the ML estimator of model $\mathcal{M}$, n is the number of observations, p is the number of parameters of model $\mathcal{M}$. The better model according to the BIC is the one presenting a lower value of B.

***Exercise 28.** *Derivate the form of the Bayesian Information Criterion, equation 4.9.*

5 Notions of probabilistic models of cognition

5.1 A simple illustration of model selection (coin flipping)

Consider the case of a coin with *bias* probability θ of giving '+' (hence, for a fair coin is $\theta = 1/2$). After N flips, the number of heads and tails (+'s and $-$'s) is N_+ and N_-, respectively, being $N = N_+ + N_-$. The likelihood of the data $d = (N_+, N_-)$ given the hypothesis θ is:

$$\mathcal{L}(d, \theta) = \theta^{N_+}(1-\theta)^{N_-} \tag{5.1}$$

the likelihood can be proved (check!) to be normalised with respect to all possible sequences of N flips.

Bayesian estimators. Supposing a uniform prior $\Pi(\theta)$ for the coin + probability, the posterior probability of having a coin with bias probability θ after a sequence of N_+ heads and N_- tails is:

$$P(\theta, d) = \frac{(N_+ + N_- + 1)!}{N_+! N_-!}\theta^{N_+}(1-\theta)^{N_-} \tag{5.2}$$

where we have used (see below) the form of the data evidence of a model with flat prior on θ

$$\mathcal{E}(d) = \int_0^\infty \mathrm{d}\theta \, \theta^{N_+}(1-\theta)^{N_-} = \frac{N_+! N_-!}{(N_+ + N_- + 1)!} \tag{5.3}$$

It is immediate to show that the MAP Bayesian estimator for θ, θ_{MAP} and the average over the posterior as a Bayesian estimator, $\hat{\theta}$ are, respectively:

$$\theta_{\mathrm{MAP}} = \frac{N_+}{N_+ + N_-} \tag{5.4}$$

$$\hat{\theta} = \frac{N_+ + 1}{N_+ + N_- + 2} \tag{5.5}$$

Model selection. Suppose that after N flips of which there are N_+ heads and N_- tails, one has to discriminate among two models: $\mathcal{M}_0$ is the model that assumes that the coin is fair (the prior distribution being equal to a delta function in $\theta = 1/2$). $\mathcal{M}_1$ is the model according to which the coin is unfair with any bias probability a priori weighted uniformly (i.e., the prior distribution is $\Pi(\theta) = 1$, as in the examples before). To discriminate among models, one may compute their respective data evidence. In the light of the previous considerations, it is straightforward to show that:

$$\mathcal{E}(d|\mathcal{M}_0) = 2^{-N} \tag{5.6}$$

$$\mathcal{E}(d|\mathcal{M}_1) = \frac{N_+! N_-!}{(N_+ + N_- + 1)!} \tag{5.7}$$

In figure 5.1-right we show the functions $\mathcal{E}(d|\mathcal{M}_0)$ and $\mathcal{E}(d|\mathcal{M}_1)$ versus N_+ for $N = 10$. In figure 5.1-left we show the posteriors $P(\theta|d, \mathcal{M}_0)$ and $P(\theta|d, \mathcal{M}_0)$ versus θ for two different $d = (N_+, N_-)$'s, corresponding to two cases in which $\mathcal{E}(d|\mathcal{M}_0) > \mathcal{E}(d|\mathcal{M}_1)$ and $\mathcal{E}(d|\mathcal{M}_0) < \mathcal{E}(d|\mathcal{M}_1)$, respectively. Although there always exist a value of θ' such that $P(\theta'|d, \mathcal{M}_1)$ explains the data better than $\mathcal{M}_0$, when taking into account the fluctuations around the MAP, for some values of $N+$, N_- the "simplest explanation", i.e., the model with lower (zero) number of parameters, $\mathcal{M}_0$, is preferred by the Bayesian model evidence.

Find a digression of this example in the context of probabilistic models of cognition in [Griffiths et al., 2008].

Exercise 29. *Consider the Dirichlet distribution of the vector of probabilities* $\mathbf{y} = (y_1, \ldots, y_n)$, $y_i \leq 0$, $\sum_{i=1}^n y_i = 1$ *with parameters* $\boldsymbol{\alpha}$*:*

$$D(\mathbf{y}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})}\prod_i y_i^{\alpha_i - 1}\, \delta\left(\sum_j y_j - 1\right) \tag{5.8}$$

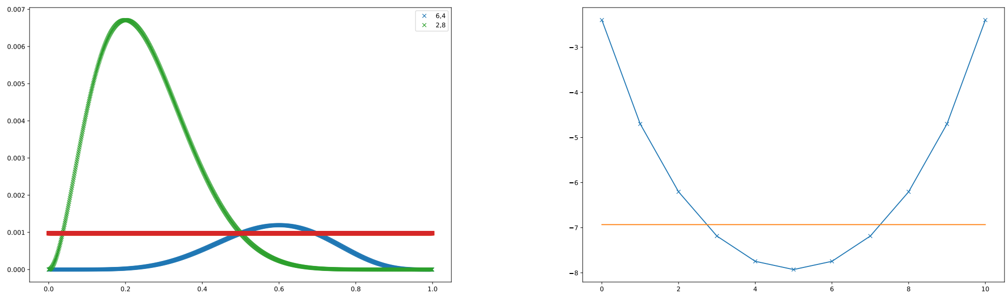

Figure 5.1: Left: $\mathcal{E}(d|\mathcal{M}_0)$ and $\mathcal{E}(d|\mathcal{M}_1)$ versus N_+ for $N = 10$. Right: $P(\theta|d, \mathcal{M}_0)$ and $P(\theta|d, \mathcal{M}_0)$ versus θ for two different $d = (N_+, N_-)$'s.

The normalisation constant is the multi-variate Beta function:

$$B(\boldsymbol{\alpha}) = \int [\mathrm{d}\mathbf{y}]\, y_j^{\alpha_j - 1} \delta\left(\sum_j y_j - 1\right) \tag{5.9}$$

where

$$[\mathrm{d}\mathbf{y}] = \prod_{j=1}^{n} \mathrm{d}y_j \tag{5.10}$$

It is possible to relate this integral to the product of Gamma functions (see, for example these notes by J. Lin, 2016):

$$B(\boldsymbol{\alpha}) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)} \tag{5.11}$$

where Γ *is the Gamma function*

$$\Gamma(\alpha) = \int_0^\infty \mathrm{d}u\, e^{-u} u^{\alpha-1}. \tag{5.12}$$

Derive the form of the data evidence and of the Bayesian estimators, eqs. 5.7, keeping in mind that $\Gamma(\alpha) = (\alpha - 1)!$*, when* α *is an integer.*

5.2 Inferring from two cues.

Suppose that the value of a quantity y (corresponding, for example, to the position of an object) generates a (stochastic) *cue* variable, c, (for example the sound that the object produces), with probability $\mathcal{L}(c|y)$. One could try to test experimentally whether a human subject is able to infer the ML value for y from the value c. The test would require the knowledge of $\mathcal{L}(c|y)$.

Suppose, alternatively, that the subject is asked to infer the value of a quantity, y, from the values of *two cues* for this quantity, c_1 and c_2. The posterior probability for y:

$$P(y|c_1, c_2) \propto \mathcal{L}(c_1, c_2|y)\pi(y) \tag{5.13}$$

We suppose that the cues do not condition one another, the likelihood of both cues is hence factorizable: $\mathcal{L}(c_1, c_2|y) = \mathcal{L}_1(c_1|y)\mathcal{L}_2(c_2|y)$. We want to put P in relation with the posterior corresponding to a single cue $P_i(y|c_i)$, $i = 1, 2$:

$$P_i(y|c_i) \propto \mathcal{L}_i(c_i|y)\pi(y) \tag{5.14}$$

It is:

$$P(y|c_1, c_2) \propto P_1(y|c_1)P_2(y|c_2)\pi(y)^{-2} \tag{5.15}$$

Let us now suppose that the single-cue posterior distributions are normal distributions $P_i(y|c_i) = \mathcal{N}(y;\mu_i(c_i),\sigma_i^2)$, and also $\pi(y) = \mathcal{N}(y;\mu,\sigma^2)$. The prediction for the optimal ML estimator, $y^*(c_1,c_2) = \arg\max_y \ln P(y|c1,c2)$ is (check it!):

$$y^*(c_1,c_2) = \mu_1(c_1)w_1 + \mu_2(c_2)w_2 - 2\mu w \tag{5.16}$$

$$w_1 = \frac{\sigma_2^2\sigma^2}{A}, \qquad w_2 = \frac{\sigma_1^2\sigma^2}{A}, \qquad w = \frac{\sigma_1^2\sigma_2^2}{A} \tag{5.17}$$

where $A = \sigma_2^2\sigma^2 + \sigma_1^2\sigma^2 - 2\sigma_1^2\sigma_2^2$.

Now, imagine that the parameters $\mu_{1,2}$ and $\sigma_{1,2}^2$, corresponding to the cognitive sensorial inference abilities of a given cognitive subject (as a human subject), can be estimated from a number of experimental estimations by the subject on the value of y given c_1 and c_2 in some dataset $\mathcal{S} = \{c_1^{(s)}\} \cup \{c_2^{(s)}\}$. One then can present *both cues at the same time* to the subject, extracted from the set $\mathcal{S}$, and see whether she infers y according to the MAP optimal solution, (5.17). See ref. [Jacobs, 1999] for a pioneer implementation of such an experiment (see also [Ernst and Banks, 2002]). Other references regarding the Bayesian approach to cognition are [Ma et al., 2006, Pouget et al., 2013, Sanborn and Chater, 2016], and refs. [Knill and Pouget, 2004, Chater et al., 2010] are reviews of the topic.

5.3 A mention to the Free Energy Principle

The Free Energy Principle for the brain [Friston, 2010, Friston et al., 2015, Schwartenbeck and Friston, 2016] is a principle for optimal perception and action of a cognitive entity. The Free Energy Principle is supported by a large corpus of cognitive and neuro-physiological experimental evidence (see the above references).

The cognitive entity is committed to the inference (*perceptual learning*) of a probabilistic generative model, or the joint probability distribution of hidden states $\mathbf{s}$ and sensorial observations $\mathbf{o}$, $p(\mathbf{s},\mathbf{o})$, where $\mathbf{s}$ and $\mathbf{o}$ are multi-dimensional vectors belonging to different spaces. The free energy principle is also a principle for optimal *learning*, or the modification of the distribution p as the subject accumulates empirical information, and for optimal *action policies* (*active inference*).

Observations $\mathbf{o}$ and (true) hidden states $\mathbf{s}$ leading to the observations (that can be in principle be assumed to coincide with those that the subject is committed to infer, $\mathbf{s}$) are generated from a joint probability distribution $R(\mathbf{s},\mathbf{o})$ called *generative process*.

Hidden states and observations are generated from the distribution $R(\mathbf{s},\mathbf{o})$. Afterwards, the subject infers $\mathbf{s}$ from the set of observations so generated. The inference of the distribution $p(\mathbf{s},\mathbf{o})$ is to be performed in such a way that the so called *surprise*, $= -\ln p(\mathbf{o})$, or minus the logarithm of the model evidence of the observations, is minimised (accounting, in this way, for the over-fitting in the case that the complexity of the function p may be modified). The minimisation of the surprise is actually done through an approximate posterior probability distribution, $q(\mathbf{s})$, approximating $p(\mathbf{s}|\mathbf{o})$. The minimisation of the surprise is done through variational free energy minimisation, in a way that is formally identical to that presented in sec. 1.3:

$$\tilde{F}[q] = KL[q(\mathbf{s})\,,\, p(\mathbf{s}|\mathbf{o})] - \ln p(\mathbf{o}) \tag{5.18}$$

in total analogy with sec. 1.3, except that now the free energy is the free energy operator of a posterior probability over hidden states, and not on a likelihood over observations, as in 1.3. The analogy is more explicitly shown in table 5.1.

Table 5.1:

Free Energy Principle	statistical physics
$\mathbf{s}$	$\mathbf{x}$
q	Q
$p(\mathbf{s}|\mathbf{o})$	$\mathcal{L}(\mathbf{x}|J)$
$p(o)$	$Z_{\mathcal{L}}$
surprise $= -\ln p(\mathbf{o})$	$F = \tilde{F}[\mathcal{L}]$

6 Notions of learning in Neural networks

Consider, as before, a phase space, Σ, with N components (or “particles”), $\mathbf{x} = (x_1, \cdots, x_N)$, and consider a set of n empirical configurations $\mathbf{x}^{(s)} \in \Sigma$, $s = 1, \cdots, n$, from which we would like to learn. “(Unsupervised) learning from them” means finding a *generative model*, or a probability distribution (a likelihood) on the $\mathbf{x}$'s, $\mathcal{L}(\mathbf{x}|W)$ with parameters W, such that the (log-)likelihood of the data is maximum. One supposes that the probability distribution exhibits an exponential form:

$$\mathcal{L}(\mathbf{x}|W) = \exp[-\mathcal{H}[\mathbf{x}, W]]/Z(W) \tag{6.1}$$

where the effective energy $\mathcal{H}$:

$$\mathcal{H}(\mathbf{x}, W) = -\mathbf{x}^\dagger W \mathbf{x} \tag{6.2}$$

and where $Z(W)$ is the normalizing constant, and W is a symmetric real matrix. This model is called Boltzmann Machine. Derivating the log-likelihood of the data with respect to the W's leads to (check!):

$$\frac{\partial \ln \mathcal{L}}{\partial W_{ij}} = n\left(x_i x_j - \langle x_i x_j \rangle_{\mathcal{L}}\right) \tag{6.3}$$

(as already seen in eq. 3.4). The W-gradient increases according to the *awake* (or experimental) correlations and decreases according to the *sleep* (or likelihood-sampled, according to the generative model) correlations. We notice again that this problem is formally equivalent to the maximum entropy inference in the presence of pairwise correlations, sec. 3.

Hidden units. Such a Boltzmann Machine is, in principle, able to efficiently capture the probability distributions of experimental ensembles that are governed by two-component (or two-particle) effective interactions. While in physics two-body correlations are often enough to efficiently describe the system, in a learning context they may be not enough[9]. To induce effective interactions between the N components of the model, N_h *hidden units* are introduced, which are additional variables of the model, to be *marginalized* (averaged out) when comparing the model with the experiment. The $N + N_h$ degrees of freedom of each configuration are now $\mathbf{x} = (\mathbf{v}, \mathbf{h})$, with $\mathbf{v} = (v_1, \cdots, v_N)$ and $\mathbf{h} = (h_1, \cdots, h_{N_h})$:

$$\ln \mathcal{L}(\mathbf{v}|W) = \ln \sum_{\mathbf{h}} \exp[\mathbf{x}^\dagger W \mathbf{x}] - \ln Z(W) \tag{6.4}$$

$$Z(W) = \sum_{\mathbf{v},\mathbf{h}} \exp[\mathbf{x}^\dagger W \mathbf{x}] \tag{6.5}$$

In this case, the gradient with respect to W of the log-likelihood of a visible configuration $\mathbf{v}$ is (check!):

$$\frac{\partial \ln \mathcal{L}(\mathbf{v}', W)}{\partial W_{ij}} = \sum_{\mathbf{h}} p_W(\mathbf{h}|\mathbf{v}')\, x'_i x'_j - \sum_{\mathbf{h},\mathbf{v}} p_W(\mathbf{v}, \mathbf{h})\, x_i x_j \tag{6.6}$$

where $p_W(\mathbf{h}|\mathbf{v}) = p_W(\mathbf{v}, \mathbf{h})/p_W(\mathbf{v})$, and $p_W(\cdot) = \mathcal{L}(\cdot|W)$, and $\mathbf{x}' = (\mathbf{v}', \mathbf{h})$.

6.1 Learning in Restricted Boltmann Machines

The Restricted Boltmann Machine is a Boltzmann machine with Boolean degrees of freedom, $x_i \in \{0, 1\}$, and hidden units such that the interaction coupling between hidden and visible variables is bipartite, i. e., $W_{ij} = 0$ for all $i, j \le N$ and for all $i, j \ge N + 1$. Re-defining the out-diagonal matrix W as an N times M matrix, w, and considering external fields, the effective energy reads:

[9] An academic example is the *shifter ensemble*, an ensemble of strings of bits such that the second half is equal to the first half, with probability 1/2, or shifted by a given number of bits to the right (with periodic boundary conditions) with probability 1/2. A Boltzmann machine without hidden units is not able to describe the ensemble with high likelihood (see [MacKay, 2003]).

$$\mathcal{H}(\mathbf{v},\mathbf{h}|W,\mathbf{b},\mathbf{c}) = -\sum_{i=1}^{N_h}\sum_{j=1}^{N} w_{ij}h_i v_j - \sum_{i=1}^{N_h} h_i c_i - \sum_{j=1}^{N} v_j b_j \tag{6.7}$$

In this circumstance, hidden variables are independent on each other, and so are visible, i. e. (we ommit the underscript in p_W): $p(\mathbf{v}|\mathbf{h}) = \prod_j p(v_j|\mathbf{h})$, and $p(\mathbf{h}|\mathbf{v}) = \prod_i p(h_i|\mathbf{v})$. On the other hand, being the probability distribution on the hidden and visible variables a separable distribution, the probability of the single visible or hidden unit to be in the state 1 is, as we saw in sec. 1.3 (where such a probability was called q_i) (chek!):

$$p(h_i = 1|\mathbf{v}) = \sigma\left(\sum_j w_{ij}v_j + c_i\right) \tag{6.8}$$

$$p(v_j = 1|\mathbf{h}) = \sigma\left(\sum_i w_{ij}h_i + b_j\right) \tag{6.9}$$

where $\sigma(y) = 1/(1+e^{-y})$. Thanks to this factorization, the first (awake) and second (sleep) terms in (6.6) are (check!):

$$\sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v})h_i v_j = p(h_i = 1|\mathbf{v})v_j \tag{6.10}$$

$$\sum_{\mathbf{v},\mathbf{h}} p(\mathbf{v},\mathbf{h})v_i h_j = \sum_{\mathbf{v}} p(\mathbf{v})p(h_i = 1|\mathbf{v})v_j \tag{6.11}$$

so that the equations for the gradient of the log-likelihood of a single state $\mathbf{x}$ are (check!):

$$\frac{\partial \ln \mathcal{L}(\mathbf{x}', W)}{\partial w_{ij}} = p(h_i = 1|\mathbf{v}')v_j' - \sum_{\mathbf{v}} p(\mathbf{v})p(h_i = 1|\mathbf{v})v_j \tag{6.12}$$

and the maximization of the likelihood of the whole training set $\mathcal{K} = \{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)}\}$:

$$\frac{\partial \ln \mathcal{L}_{\mathrm{all}}(\mathcal{K}, W)}{\partial w_{ij}} = \frac{1}{n}\sum_{s=1}^{n} \langle v_j^{(s)} h_i\rangle_{p(h_i|\mathbf{v}^{(s)})} - \langle h_i v_j\rangle_{p(\mathbf{v},\mathbf{h})} = \tag{6.13}$$

$$= \frac{1}{n}\sum_{s=1}^{n} p(h_i = 1|\mathbf{v}^{(s)})v_j^{(s)} - \sum_{\mathbf{v}} p(\mathbf{v})p(h_i = 1|\mathbf{v})v_j \tag{6.14}$$

(6.15)

(where $\mathcal{L}_{\mathrm{all}}$ is the joint probability distribution of all the $\mathbf{v}$'s) the log-likelihood derivative w.r.t. the fields are (check!):

$$\frac{\partial \ln \mathcal{L}(\mathbf{x}', W)}{\partial b_j} = v_j' - \sum_{\mathbf{v}} p(\mathbf{v})v_j \tag{6.16}$$

$$\frac{\partial \ln \mathcal{L}(\mathbf{x}', W)}{\partial c_i} = p(h_1 = 1|\mathbf{v}') - \sum_{\mathbf{v}} p(\mathbf{v})p(h_1 = 1|\mathbf{v}) \tag{6.17}$$

In this way, one has reduced the complexity of the problem to the computation of only one ensemble average, the one corresponding to the sleep term. This ensemble calculation (the second term in (6.14)) can be performed in a particularly simple way due to the fact that, given the $\mathbf{v}$'s, the $\mathbf{h}$'s can be sampled (in parallel, by the way), and vice-versa, with the help of a Gibbs Monte-Carlo algorithm, in the following way:

1. propose a vector $\mathbf{v}(0)$ (for example $= \mathbf{v}^{(s)}$ for $s = 1, \ldots, n$)
2. for $t = 0, \ldots, T$:

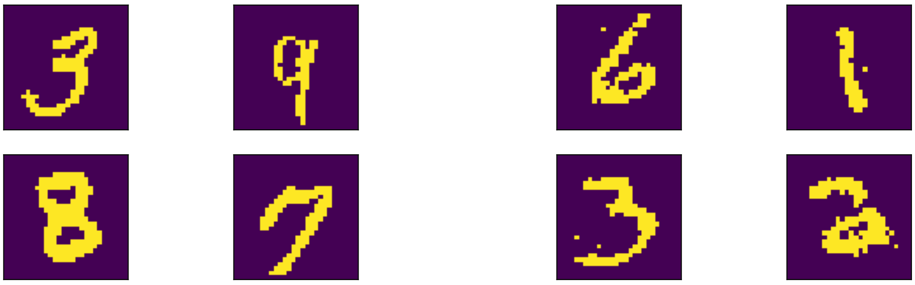

Figure 6.1: Four letters belonging to the training dataset of the MNIST ensemble (left), and four letters generated (with a *daydream* Gibbs algorithm) from the inferred likelihood $\mathcal{L}_{\text{all}}$ (right). The learned letters seem handwritten.

   - sample a vector $\mathbf{h}(t) \sim p(h_i|\mathbf{v}(t))$
   - sample a vector $\mathbf{v}(t+1) \sim p(v_i|\mathbf{h}(t))$

3. approximate the sleep term in (6.14), $\sum_{\mathbf{v}} p(\mathbf{v})p(h_i = 1|\mathbf{v})v_j$, by: $p(h_i = 1|\mathbf{v}(T))v_j(T)$

This algorithm is called the T-step *contrastive diverence* algorithm [Fischer and Igel, 2012].

The Maximum Likelihood algorithm for the updating of the W's is, finally:

1. propose an initial value of the couplings, $W(0)$
2. for $\mathsf{r} = 0, \ldots, \mathsf{R}$:
   (a) compute (6.14), approximating the sleep term with the T-step contrastive divergence agorithm, schetched before, and call it δW_{ij}
   (b) as in (2.11),
   $$W_{ij}(\mathsf{r}+1) = W_{ij}(\mathsf{r}) + \eta\, \delta W_{ij}(\mathsf{r}) \tag{6.18}$$
   the η parameter is called the *learning rate*. The term $\delta W_{ij}(\mathsf{r})$ is equal to equation 6.14. The second (*sleep*) term contains a sum over all possible values of $\mathbf{v}$, weighted with $p(\mathbf{v})$. Such a sampling is approximated with the contrastive divergence algorithm described above.

6.2 Two examples of unsupervised learning in RBM's

As an illustration of the learning process in a RBM with the above described algorithm, we present two examples of unsupervised learning in RBM with the contrastive divergence algorithm.

The first example is the learning of the MNIST database [MNI,] of hadwritten digits. We have learned $K = 10^4$ binary MNIST samples of resolution 28×28 (flattened) handritten digits (fig. 6.1, left), with parameters $M = 28^2$ $N = M/2$, $\eta = 0.02$. After $\mathsf{R} = 5 \cdot 10^4$ iterations, the RBM succeeds in learning with relatively high likelihood more than half of the digits of a test dataset (different from the K digits used in the training set). The learned digits have been learned with hight likelihood. In fig. 6.1, right we show four random letters extracted from the learned $\mathcal{L}_{\text{all}}$, $\mathbf{v} \sim \mathcal{L}_{\text{all}}(\cdot|W^*)$.

The second example is learning in the shifter ensemble (SE) (c. f. footnote 9). We define the (n, m)-SE as the ensemble of strings of n bits in such a way that the second half of the string (n is even) is either equal to the first half (with probability 1/3), or shifted by m positions (with probability 1/3), or shifted by $-m$ positions (i. e., by m positions at left, with probability 1/3), with periodic boundary conditions in the second half. For example, both 010100 and 010010 belong to the $(6, 1)$-SE. As we mentioned, such an ensemble is not learnable with (pairwise-correlation) maximum entropy. The RBM has learned a training set of $K = 10^4$ random instances of the $(24, 1)$-SE, with parameters $M = 24$ $N = M$, $\eta = 0.02$. After $\mathsf{R} = 2 \cdot 10^3$ iterations, the RBM succeeds in learning with relatively high likelihood roughly the 90% of the test dataset.

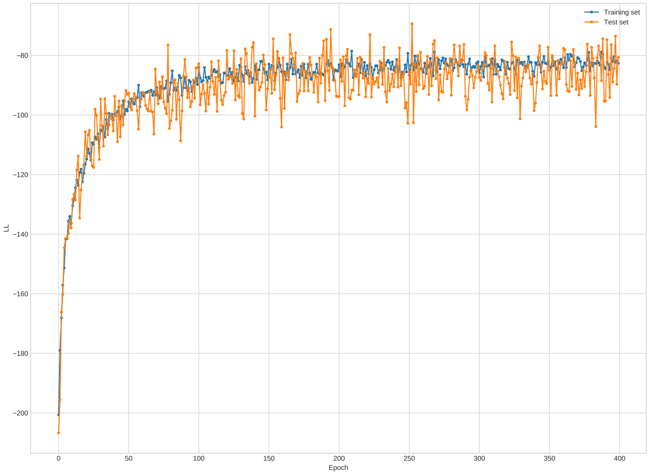

Figure 6.2: Log-likelihood of the training and test (MNIST) datasets as a function of the number of iterations r. The learning has parameters: $n = 4 \cdot 10^4$, $n_{\text{test}} = 2 \cdot 10^4$, $\eta = 0.02$, $M = 784 = 32^2$, $N = M/12$, $T = 1$ (batch-size $B = 50$).

A Appendix: the Metropolis algorithm

Metropolis-Hastings algorithm. A general way of constructing a Markov Chain is first proposing a transition from state i-th to j-th defined by the *proposal matrix* $p_{ij}^{(0)}$ (where $p^{(0)}$ is a stochastic irreducible matrix), and accepting it with probability a_{ij}. The transition matrix is hence $p_{ij} = a_{ij} p_{ij}^{(0)}$ for $i \neq j$ and $p_{ii} = p_{ii}^{(0)} + \sum_{j \neq i} p_{ij}^{(0)} (1 - a_{ij})$ (for the correct normalization it is necessary to keep refused configurations). Detailed balance is satisfied if

$$a_{ij} = F\left(\frac{\pi_j p_{ji}^{(0)}}{\pi_i p_{ij}^{(0)}}\right) \tag{A.1}$$

being $F : \mathbb{R}^+ \to [0:1]$ satisfying $F(x) = xF(1/x)$. The *Metropolis algorithm* corresponds to:

$$F(x) = \min\{x, 1\} \tag{A.2}$$

If the proposal matrix satisfies detailed balance, all the proposals are accepted. For any symmetric irreducible proposal matrix, the acceptance probabilities depends only on the ratio between the target distribution probabilities:

$$a_{ij} = \min\left\{\frac{\pi_j}{\pi_i}, 1\right\} \tag{A.3}$$

in the canonical ensemble at inverse temperature β, for example, this reads to $a_{ij} = \min\{1, \exp(-\beta N(\epsilon_i - \epsilon_j))\}$ where ϵ_j are the per site energy of the j-th configuration.

Single-particle updating. Consider a system composed by N degrees of freedom $\mathbf{s} = \otimes_{m=1}^{N} \sigma^{(m)}$, where $\sigma^{(m)}$ is the m-th particle state. Let $p^{(m)}$ be the transition matrix in which only particle m is updated:

$$p_{ij}^{(m)} > 0 \qquad \sigma_i^{(n)} = \sigma_j^{(n)} \qquad \forall n \neq m \tag{A.4}$$

$$p_{ij}^{(m)} = 0 \qquad \text{otherwise} \tag{A.5}$$

Updating a random sequence of particles, one at once, is called *random-particle updating*, and a sequence of N random particle updating is called a *sweep*, the corresponding transition matrix being $p = (1/N) \sum_{m=1}^{N} p^{(m)}$. If the particles are updated following a given sequence of indices $i_1, \ldots, i_N$, the updating is called *sequential*, the corresponding transition matrix being $p = \prod_{m=1}^{N} p^{(i_m)}$. Still a different scheme is called *M- multi-hit algorithm*, in which one selects one particle, and applies the Metropolis algorithm M times (proposing a new state for particle m and accepting it with matrix a), whose transition matrix corresponds to $p = (1/N) \sum_{m=1}^{N} [p^{(m)}]^M$. If the single-particle transition matrices satisfy detailed balance, so does the random-particle updating matrix, while the sequential matrices satisfy, in general, only the balance condition (which is the required condition for a valid MC).

A.1 A Metropolis algorithm for the likelihood estimation of the Gaussian mixture inference problem

For the case of the Gaussian mixture, a valid Metropolis algorithm reads: one choses random initial conditions $\boldsymbol{\theta}^{(0)}$, then:

1. at the t-th iteration, one performs an attempt $\tilde{\mu}_j = \mu_j^{(t)} + \xi$ where $\xi \sim \mathcal{N}(0, \eta)$ being η a parameter (to be optimized). The constraint parameters $p_j^{(t)}$ can be updated as $\ln \tilde{p}_j = \ln p_j^{(t-1)} + \zeta$ being $\zeta \sim \mathcal{N}(0, \eta^2)$ (see Exercise **??**), eventually evaluating this trial with a further prior probability $\pi(\tilde{\boldsymbol{\theta}})$.
2. Application of the Metropolis rule: with probability: $r = f(\mathbf{x}|\tilde{\boldsymbol{\theta}})\pi(\tilde{\boldsymbol{\theta}}) / [f(\mathbf{x}|\boldsymbol{\theta}^{(t)})\pi(\boldsymbol{\theta}^{(t)})]$ accept the trial, $\boldsymbol{\theta}^{(t+1)} \equiv \tilde{\boldsymbol{\theta}}$; $t++$, go to 1.

B Appendix: cumulant expansion

B.1 One-dimensional cumulant expansion

Let x be a random variable and $\langle\cdot\rangle$ its average according to a give distribution P. The logarithm of the *generating function*:

$$g(u) = \langle e^{ux} \rangle \tag{B.1}$$

can be expanded as a series in cumulants κ_n:

$$\ln g(u) = \sum_{n=1}^{\infty} \frac{u^n}{n!} \kappa_n \tag{B.2}$$

κ_n is a function of the first n moments of the P distribtion. In particular, the first one-dimensional cumulants are, in terms of the moments $\mu_n = \langle x^n \rangle$:

$$\begin{aligned}
\kappa_1 &= \mu_1 && \text{(B.3)}\\
\kappa_2 &= \mu_2 - \mu_1^2 && \text{(B.4)}\\
\kappa_3 &= \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3 && \text{(B.5)}\\
\kappa_4 &= \mu_4 - 4\mu_2^2\mu_1\mu_3 + 12\mu_2\mu_1^2 - 6\mu_1^4 && \text{(B.6)}\\
& && \text{(B.7)}
\end{aligned}$$

(the functional relation in terms of central moments $\langle (x - \langle x \rangle)^n \rangle$ is identical taking $\mu_1 = 0$). The moments emerge from the generating function, which is the function generating the moments (while its logarithm is the function generating the cumulants):

$$g(u) = \sum_{n=0}^{\infty} \frac{u^n}{n!} \mu_n. \tag{B.8}$$

B.2 Multivariate cumulant expansion

Consider a multivariate distribution P on n real variables. One defines the generating function

$$g(\mathbf{u}) = \langle e^{\mathbf{u}\cdot\mathbf{x}} \rangle \tag{B.9}$$

where $\langle\cdot\rangle = \int [\mathrm{d}\mathbf{x}] \cdot P(\mathbf{x})$. The multivariate cumulant expansion is:

$$\ln g(\mathbf{u}) = \sum_{p=1}^{\infty} \sum_{i_1,\cdots,i_p} \frac{1}{p!} u_{i_1} \cdots u_{i_p} \, \kappa(i_1, \cdots, i_p) \tag{B.10}$$

where $\kappa(i_1, \cdots, i_p)$ is the joint cumulant of variables $x_{i_1}, \cdots, x_{i_p}$. It is:

$$\kappa(i_1, \cdots, i_p) = \sum_{\pi \in \mathcal{P}} (-1)^{|\pi|-1} (|\pi| - 1)! \prod_{B \in \pi} \langle \prod_{j \in B} x_j \rangle \tag{B.11}$$

where $\mathcal{P}$ is the set of all partitions of the ensemble of p indices $(i_1, \cdots, i_p)$; $B \in \pi$ is the block of variables in the partition π, and $|\pi|$ is the number of blocks in the partition. For example, if the set is 1234, there are: 1× [4]; 6×[1 + 1 + 2]; 4 × [1 + 3]; 3 × [2 + 2] and 1 × [1 + 1 + 1 + 1] partitions. The value of $|\pi|$ for them is 1, 3, 2, 2, 4, respectively. For example, supposing null averages, the cumulant $\kappa(1, 2, 3, 4, 5, 6)$ is

$$\begin{aligned}
\kappa(1,2,3,4,5,6) &= \overline{123456} - \\
&- \overline{1234} \cdot \overline{56} + \text{all other } [2+4] \text{ combinations} + \\
&+ 2\left(\overline{12} \cdot \overline{34} \cdot \overline{56} + \text{all other } [2+2+2] \text{ combinations}\right)
\end{aligned} \tag{B.12}$$

where $\overline{1234}$ is a shorthand for $\langle x_1 x_2 x_3 x_4 \rangle$.

The $i_1, \cdots, i_n$ cumulant is also called n-body connected correlator, also denoted by $\langle\langle x_{i_1} \cdots x_{i_n} \rangle\rangle$. It can be obtained by differentiating the log-generating function:

$$\kappa(i_1, \cdots, i_n) = \frac{\partial^n}{\partial_{u_{i_1}} \cdots \partial_{u_{i_n}}}\Bigg|_{\mathbf{u}=\mathbf{0}} \ln g(\mathbf{u}) \tag{B.13}$$

C The Gaussian Model and the Wick's theorem

Let us define the multi-dimensional Gaussian probability distribution defined on the real valued N-dimensional vectors $\mathbf{x} \in \mathbb{R}^N$:

$$P(\mathbf{x}) = \frac{1}{Z} \exp[-H(\mathbf{x})] \tag{C.1}$$

$$H(\mathbf{x}) = \frac{1}{2} \sum_{k,l=1}^{N} x_k J_{kl} x_l \tag{C.2}$$

where J is a real, symmetric, positively definite matrix (with rank $= N$), and Z is the normalising constant.

Connected correlation functions (cumulants) of any order. We would like to calculate the correlation functions at any (even) order, n. These are defined by

$$\langle\langle x_{s_1} x_{s_2} ... x_{s_n} \rangle\rangle = \frac{1}{Z} \int_{-\infty}^{\infty} [\prod_{k=1}^{N} dx_k] x_{s_1} x_{s_2} ... x_{s_n} e^{-\frac{1}{2} \sum_{k,l=1}^{N} x_k J_{kl} x_l} \tag{C.3}$$

In order to compute them, we define a *generating functional*:

$$Z[\mathbf{h}] = \int_{-\infty}^{+\infty} [\prod_{k=1}^{N} dx_k] e^{-\frac{1}{2} \sum_{k,l=1}^{N} x_k J_{kl} x_l + \sum_{m=1}^{N} h_m x_m} \tag{C.4}$$

Were $\mathbf{h} = (h_1, \ldots, h_N)$. A useful property of the generating functional (see appendix B) is that the connected correlations functions are obtained from its derivatives with respect to the fields. Indeed, an arbitrary correlation function of n different x's is given by

$$\langle\langle x_{s_1} x_{s_2} ... x_{s_n} \rangle\rangle = \frac{1}{Z} \frac{\partial^n Z[\mathbf{h}']}{\partial h'_{s_1} \partial h'_{s_2} ... \partial h'_{s_n}} |_{\mathbf{h}=\mathbf{0}} \tag{C.5}$$

Here $Z \equiv Z[0]$ is the partition function. Let us calculate $Z[\mathbf{h}]$. Under the (unitary) change of variables:

$$\mathbf{y} = \mathbf{x} - J^{-1}\mathbf{h} \tag{C.6}$$

the integral for $Z[\mathbf{h}]$ becomes:

$$Z[\mathbf{h}] = \int_{-\infty}^{+\infty} [\prod_{k=1}^{N} dx_k] e^{-\frac{1}{2} \sum_{k,l=1}^{N} y_k J_{kl} y_l + \frac{1}{2} \sum_{k,l=1}^{N} h_k (J^{-1})_{kl} h_l} = \tag{C.7}$$

$$= \frac{(2\pi)^{N/2}}{\sqrt{\det J}} e^{\frac{1}{2} \sum_{k,l=1}^{N} h_k (J^{-1})_{kl} h_l} \tag{C.8}$$

where we have used the Gaussian integration rule [10].

Partition Function The partition function has a simple expression, that can be obtained either imposing $\mathbf{h} = \mathbf{0}$ in C.8, or computing it from scratch:

$$Z = \int [\prod_{k=1}^{N} dx_k] e^{-\frac{1}{2} \sum_{k,l} J_{kl} x_k x_l} = \int [\prod_{k=1}^{N} dx_k] e^{-\frac{1}{2} \sum_k x_k'^2 \epsilon_k}$$

$$= \frac{(2\pi)^{N/2}}{(\det J)^{1/2}} \tag{C.9}$$

where ϵ_k are the J eigenvalues, and $\mathbf{x}' = E\mathbf{x}$ are the coordinates $\mathbf{x}$ in the basis of the eigenvectors of J, or: $EJE^\dagger = \text{diag}(\epsilon_1, \ldots, \epsilon_N)$.

[10] $\int_{-\infty}^{\infty} \mathrm{d}x \exp(-Ax^2 + Bx) = (\pi/A)^{1/2} e^{B^2/4A}$.

Two-point correlations Now we calculate the $n = 2$-point correlations. Particularising equations C.5 and C.8 for $n = 2$ one obtains:

$$\langle\langle x_k x_m \rangle\rangle = \langle x_k x_m \rangle = (J^{-1})_{km} \tag{C.10}$$

We have derived in this way a fundamental and useful result: the two-point correlation matrix of the Gaussian model results to be the inverse of the interaction matrix.

Wick's theorem Wick's theorem relates the evaluation of many-point correlation functions with a function of different two-point correlations (see, for example, [?]). In the present context, it directly follows from the evaluations of the derivatives in equation C.5. It states that

$$\langle x_{s_1} x_{s_2} ... x_{s_n} \rangle = \langle x_{s_1} x_{s_2} \rangle \langle x_{s_3} x_{s_4} \rangle \cdots \langle x_{s_{n-1}} x_{s_n} \rangle + \cdots \tag{C.11}$$

where the dots after the + stand for all other $(n-1)!/(2^{n/2-1}\,(n/2-1)!)$ possible pairings of x coordinates into pairs $x_i x_j$. For example,

$$\langle x_{s_1} x_{s_2} x_{s_3} x_{s_4} \rangle = \langle x_{s_1} x_{s_2} \rangle \langle x_{s_3} x_{s_4} \rangle + \langle x_{s_1} x_{s_3} \rangle \langle x_{s_2} x_{s_4} \rangle + \langle x_{s_1} x_{s_4} \rangle \langle x_{s_2} x_{s_3} \rangle \tag{C.12}$$

Since the averages of the single variables vanish, $\langle x_i \rangle = 0$, the correlation functions of and odd number n of terms vanish as well, as it can be immediately derived from equation C.3.

This implies (see (B.11)) that all $p > 2$-cumulants (or connected correlators) are zero in the Gaussian model.

References

[MNI,] Mnist database. http://yann.lecun.com/exdb/mnist/.

[Amit and Martin-Mayor, 2005] Amit, D. J. and Martin-Mayor, V. (2005). *Field theory, the renormalization group, and critical phenomena*. McGraw-Hill International Book Co.

[Anderson Jr and Morley, 1985] Anderson Jr, W. N. and Morley, T. D. (1985). Eigenvalues of the laplacian of a graph. *Linear and multilinear algebra*, 18(2):141–145.

[Bengio, 2009] Bengio, Y. (2009). Learning deep architectures for ai. *Foundations and Trends in Machine Learning*, 2(1):1–127.

[Berg, 2017] Berg, J. (2017). Statistical mechanics of the inverse ising problem and the optimal objective function. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(8):083402.

[Bethge and Berens, 2008] Bethge, M. and Berens, P. (2008). Near-maximum entropy models for binary neural representations of natural images. In *Advances in neural information processing systems*, pages 97–104.

[Bialek et al., 2012] Bialek, W., Cavagna, A., Giardina, I., Mora, T., Silvestri, E., Viale, M., and Walczak, A. M. (2012). Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences*, 109(13):4786–4791.

[Cavagna et al., 2015] Cavagna, A., Del Castello, L., Dey, S., Giardina, I., Melillo, S., Parisi, L., and Viale, M. (2015). Short-range interactions versus long-range correlations in bird flocks. *Phys. Rev. E*, 92:012705.

[Cavagna et al., 2017] Cavagna, A., Giradina, I., Skert, C., and Viale, M. (2017). *(work in progress)*.

[Chater et al., 2010] Chater, N., Oaksford, M., Hahn, U., and Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6):811–823.

[Ernst and Banks, 2002] Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429.

[Fischer and Igel, 2012] Fischer, A. and Igel, C. (2012). *An Introduction to Restricted Boltzmann Machines*, pages 14–36. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Friston, 2010] Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127.

[Friston et al., 2015] Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive neuroscience*, 6(4):187–214.

[Griffiths et al., 2008] Griffiths, T. L., Kemp, C., and Tenenbaum, J. B. (2008). Bayesian models of cognition. In *Annual Meeting of the Cognitive Science Society, 2004; This chapter is based in part on tutorials given by the authors at the aforementioned conference as well as the one held in 2006.* Cambridge University Press.

[Ibanez-Berganza, 2016] Ibanez-Berganza, M. (2016). Monte carlo methods in statistical physics. http://www.fis.unipr.it/home/miguel.berganza/montecarlo2016.html.

[Ibanez-Berganza et al., 2018] Ibanez-Berganza, M., Amico, A., and Loreto, V. (2018). *(work in progress)*.

[Ibáñez-Berganza et al., 2019] Ibáñez-Berganza, M., Lancia, G. L., Amico, A., Monechi, B., and Loreto, V. (2019). Unsupervised inference approach to facial attractiveness. *arXiv:1910.14072*.

[Jacobs, 1999] Jacobs, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision Research*, 39(21):3621 – 3629.

[Jaynes, 1957a] Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630.

[Jaynes, 1957b] Jaynes, E. T. (1957b). Information theory and statistical mechanics. ii. *Phys. Rev.*, 108:171–190.

[Kappen and Rodríguez, 1998] Kappen, H. J. and Rodríguez, F. d. B. (1998). Efficient learning in boltzmann machines using linear response theory. *Neural Computation*, 10(5):1137–1156.

[Knill and Pouget, 2004] Knill, D. C. and Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712 – 719.

[Ma et al., 2006] Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432.

[MacKay, 2003] MacKay, D. J. (2003). *Information theory, inference and learning algorithms.* Cambridge university press.

[Marinari and Parisi, 2004] Marinari, E. and Parisi, G. (2004). Trattatello di probabilità. http://www.phys.uniroma1.it/DipWeb/web_disp/d3/dispense/marinari-parisi–prob.pdf.

[Morcos et al., 2011] Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301.

[Mussardo, 2010] Mussardo, G. (2010). *Statistical field theory: an introduction to exactly solved models in statistical physics.* Oxford University Press.

[Nguyen et al., 2017] Nguyen, H. C., Zecchina, R., and Berg, J. (2017). Inverse statistical problems: from the inverse ising problem to data science. *Advances in Physics*, 66(3):197–261.

[Pouget et al., 2013] Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature neuroscience*, 16(9):1170.

[Roudi et al., 2009] Roudi, Y., Aurell, E., and Hertz, J. A. (2009). Statistical physics of pairwise probability models. *Frontiers in computational neuroscience*, 3:22.

[Sakellariou et al., 2016] Sakellariou, J., Tria, F., Loreto, V., and Pachet, F. (2016). Maximum entropy models capture melodic styles. *arXiv preprint arXiv:1610.03414*.

[Sanborn and Chater, 2016] Sanborn, A. N. and Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, 20(12):883 – 893.

[Schneidman et al., 2006] Schneidman, E., Berry, M. J., Segev, R., and Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012.

[Schwartenbeck and Friston, 2016] Schwartenbeck, P. and Friston, K. (2016). Computational phenotyping in psychiatry: A worked example. *eNeuro*, 3(4).

[Shlens et al., 2009] Shlens, J., Field, G. D., Gauthier, J. L., Greschner, M., Sher, A., Litke, A. M., and Chichilnisky, E. J. (2009). The structure of large-scale synchronized firing in primate retina. *Journal of Neuroscience*, 29(15):5022–5031.

[Shlens et al., 2006] Shlens, J., Field, G. D., Gauthier, J. L., Grivich, M. I., Petrusca, D., Sher, A., Litke, A. M., and Chichilnisky, E. J. (2006). The structure of multi-neuron firing patterns in primate retina. *Journal of Neuroscience*, 26(32):8254–8266.

[Tang et al., 2008] Tang, A., Jackson, D., Hobbs, J., Chen, W., Smith, J. L., Patel, H., Prieto, A., Petrusca, D., Grivich, M. I., Sher, A., Hottowy, P., Dabrowski, W., Litke, A. M., and Beggs, J. M. (2008). A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *Journal of Neuroscience*, 28(2):505–518.