Inference and learning

A statistical physics perspective

miguel.berganza@roma1.infn.it

Contents

Abstract

Compressed lecture notes of a three-lesson seminar about inference and learning. We describe maximum entropy inference and learning in Restricted Boltzmann Machines, with emphasys on the analogies between both[1].

1 Introduction: the direct problem in three examples

By *direct problem* one means to find a finite number of *particular* instances generated with a *general* rule.

[1] Part of the course *Sistemi Complessi*, Laurea Magistrale in Fisica, Università di Roma, "La Sapienza", anno accademico 2017-2018.

1.1 Sorting random numbers according to a given distribution

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 distributed according to f.

Exercise 1. *What about a distribution of which one does not known the analytical form for its cumulative? Develop an algorithm generating a couple of normally distributed variables in polar coordinates using the cumulative method (Box-Mueller algorithm, 1985).*

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

Consider a discrete space $\mathbf{\Sigma}$ of $\mathcal{N}$ configurations (or states, $\boldsymbol{\sigma}_i$, $i = 1, \ldots, \mathcal{N}$)[2].

Consider a (non-factorisable) target probability distribution $\boldsymbol{\pi} = (\pi_i)_{i=1}^{\mathcal{N}}$, $\pi_i = \text{prob}(\boldsymbol{\sigma}_i)$ for $i = 1, \ldots, \mathcal{N}$. The probability is non-factorisable in what it cannot be expressed as a product of the various degrees of freedom of the configuration. In other words, two-degrees of freedom correlations are nonzero according to $\boldsymbol{\pi}$ (think, for example, in a Boltzmannian probability distribution in a given ensemble, with energy given by a pairwise interaction Hamiltonian). One cannot simply use the previous method of sorting configurations.

Example 1. *Inefficiency of uniform sampling MC 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 $\tilde{\Phi} = N(\epsilon - Ts)$ is the free energy functional (and s is the microcanonical entropy), N is the system mass, and Z is the partition function. In saddle-point approximation, it is $Z_\beta = \exp[-N\beta\Phi(\beta)]$, where $\Phi(\beta) = \min_\epsilon \tilde{\Phi}(\epsilon,\beta)$ is the 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 a random configuration (as those sampled in an unbiased MC) has exponentially suppressed probability of not having ϵ_0. On their turn, they have exponentially vanishing probability in an ensemble at $\beta > 0$.*

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 p whose element p_{ij} is the transition probability of the i-th to the j-th state, $i, j = 1, \ldots, \mathcal{N}$. The transition matrix is a stochastic matrix, it satisfies: $p_{ij} \geq 0 \, \forall i, j$ and $\sum_j p_{ij} = 1$. The Markov Chain characterized by p is said *irreducible* if given any two states i, j, one can reach j form i in a finite time, i.e., if there exists n such that $(p^n)_{ij} > 0$. A stronger property is *aperiodicity*: if there exists a n such that $(p^t)_{ij} > 0$ for all i, j, and for all $t > n$.

The matrix p 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, $\boldsymbol{\sigma}_{i_1}\boldsymbol{\sigma}_{i_2}\ldots$, which is $\pi^{(0)}(\boldsymbol{\sigma}_{i_1})\, p_{i_1 i_2} p_{i_2 i_3} \ldots$, and the probability of having the j-th state at time t is $= \sum_i (p^t)_{ij} \pi^{(0)}(\boldsymbol{\sigma}_i)$.

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

1. *The limit $\pi_j = \lim_{n\to\infty}(p^n)_{ij}$ uniquely exists, independently on i. $\pi_j \equiv \pi(\boldsymbol{\sigma}_j)$ is a PD ($\sum_j \pi_j = 1$), stationary under p:*

$$\pi_j = \sum_i p_{ij}\pi_i \qquad \text{Balance condition} \tag{1.1}$$

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

[2] we will deal with states composed by N degrees of freedom, $\mathbf{\Sigma} = \mathbf{\Sigma}_1^{\otimes N}$, where $\mathbf{\Sigma}_1$ is the single-particle degree of freedom (a binary spin $\mathbf{\Sigma}_1 = \{0,1\}$, in the case of the Ising model, for example), each one with D degrees of freedom, and $\mathcal{N} = D^N$

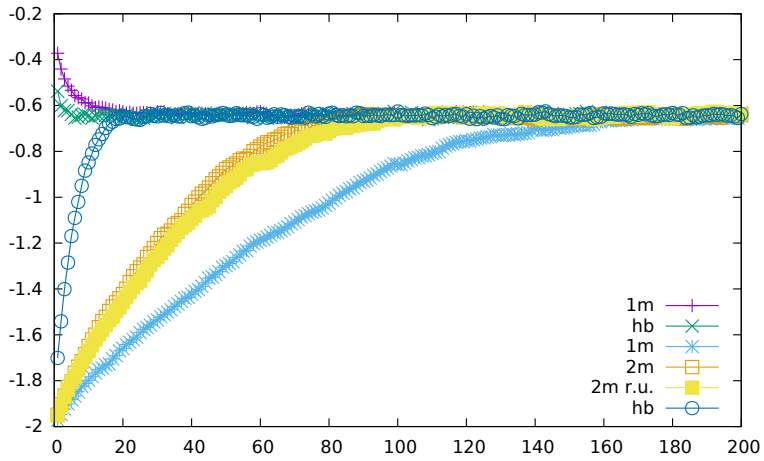

Figure 1.1: Energy vs. number of MC sweeps for the $q = 10$ 2D Potts model in the square lattice with periodic boundary conditions at $\beta = 1.24$, using several MC algorithms (Metropolis, heatbath, 2-hit Metropolis, 2-hit with random sweeps), and starting from ordered and disordered configurations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

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n} f(\boldsymbol{\sigma}^{(t)}) = \sum_{i=1}^{\mathcal{N}} \pi_i f_i \tag{1.2}$$

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

Exercise 2. *Let us define the distance between two distributions $||\boldsymbol{\alpha}-\boldsymbol{\beta}|| = \sum_j |\alpha_j - \beta_j|$. 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 $\boldsymbol{v}^\dagger p$ and $\boldsymbol{\pi}^\dagger$ (use the triangle inequality). This proves that the stationary distribution is a fixed point of the matrix p.*

The dynamic (or Markov-Chain) Monte Carlo method consists in choosing a transition matrix P such that its stationary distribution π is the desired one. The theorem before requires for the dynamic MC method to work, that 1) P must be irreducible and 2) that it satisfies the Balance condition.

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 Hamiltonian of the configuration is plotted as a function of the number of samples. Various updating algorithms are compared (among which the single-spin Gibbs sampling update, with sequential and random updatings).

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:

$$\mathcal{L}(\mathbf{x}|\beta, J) = Z(\beta, J)^{-1} \exp\left[-\beta E[\mathbf{x}, J]\right] \tag{1.3}$$
$$\tag{1.4}$$

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 distribution Q (considered to be generic):

$$\beta\tilde{F}[Q, \beta, J] = \beta\langle E[\mathbf{x}, J]\rangle_Q - S[Q] = \tag{1.5}$$
$$= \langle \ln \frac{Q}{\mathcal{L}}\rangle_Q - \ln Z(\beta, J) = \tag{1.6}$$
$$= \beta F(\beta, J) + \mathrm{KL}[Q, \mathcal{L}] = \beta\tilde{F}[\mathcal{L}, \beta, J] + \mathrm{KL}[Q, \mathcal{L}] \tag{1.7}$$

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).

We now particularize for the Ising model. The configuration space is $\mathbf{x}_i \in \{-1, 1\}$, and the energy

$$E[\mathbf{x}, J] = -\mathbf{x}^\dagger J\mathbf{x} - \mathbf{h}^\dagger\mathbf{x} \tag{1.8}$$

where, J is a real symmetric matrix with zero diagonal, $\mathbf{h}$ is a real vector (the dependence of E in $\mathbf{h}$ is absorbed into that of J).

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

$$Q(\mathbf{x}, \mathbf{a}) = \frac{e^{\sum_{i=1}^n x_i a_i}}{Z_Q}. \tag{1.9}$$

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:

$$\langle E[\mathbf{x}, J]\rangle_Q = -\sum_{i,j=1}^n \frac{1}{2} J_{ij}\bar{x}_i\bar{x}_j - \sum_{i=1}^n h_i\bar{x}_i \tag{1.10}$$

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

Derivating the variational free energy, (1.5), with respect to a_i and equating to zero leads (check!) to the set of coupled equations:

$$\begin{aligned} a_i &= \beta\,[\sum_{j=1}^n J_{ij}\bar{x}_j + h_i] \\ \bar{x}_j &= \tanh a_j \end{aligned}. \tag{1.11}$$

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

$$a = \beta\ (CJ\bar{x} + h) \qquad \bar{x} = \tanh a \tag{1.12}$$

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.

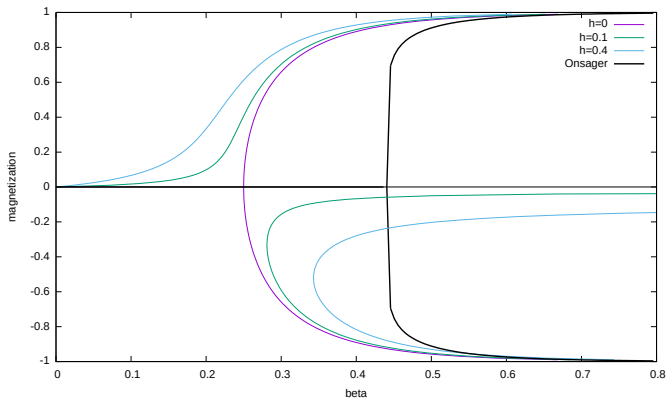

Figure 1.2: $\bar{m}(\beta)$.

Exercise 3. *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 identity* $\ln x \leq x - 1$*).*

2 Inference: basic notions and two 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:

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

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[3], 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 4. *Obtain the marginal probability distribution for the average, $\mathcal{L}(D|\mu)$.*

[3] 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.

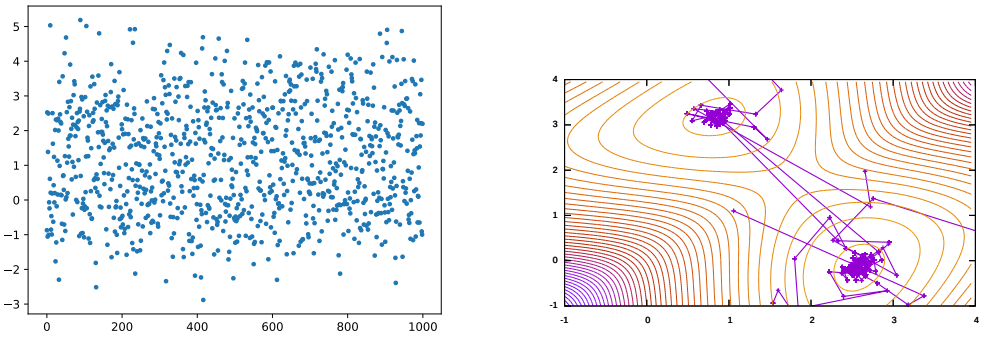

Figure 2.1: Successive values of the parameters $\theta_{1,2}$ found by 16 realizations of the EM algorithm (left) and of the Metropolis algorithm (right), for a mixture of 2 Gaussians, and a likelihood (represented by contour iso-likelihood lines) corresponding to $n = 500$, points (see the details, the algorithm scripts and the data in /BayesianMixture/Metropolis/). The probabilities are supposed to be known. The EM algorithm converges to each of the two maxima, depending on the initial conditions (only the absolute maximum corresponds to (but do not coincide with) the true parameters used to generate the data, $\mu_1 = 2.5$, $\mu_2 = 0$). The Metropolis algorithm samples both maxima in every run but, asymptotically, the absolute maxima is infinitely more sampled.

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):

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

We will consider the simple case of a mixtures of Gaussians: $\mathcal{N}(\theta_i, 1)$, i.e., $f(x_i|\theta_j) = (2\pi)^{-1/2} \exp(-(x_i - \theta_j)^2/2)$.

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 5). One chooses an algorithm (see the Appendix 5) 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 prooved 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. (1.2).

Numerical example Let us show a numerical example. We will consider a 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, $f = \mathcal{N}$. We generate a set of $n = 500$ points with known (to be inferred a posteriori) averages $\mu_1 = 2.5$, $\mu_2 = 0$.

Fig. 2.1 shows one hundred points generated with $\mathcal{L}(\mathbf{x}|\boldsymbol{\theta})$ (right). 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.

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 $\mathbf{x} = \{x_i\}_{i=1}^n$. Suppose that one has M 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 $O_k^{(i)}$ is the i-th experimental result (a number), $i = 1, \cdots, M$, of the k-th observable.

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:

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

In other words the *maximum entropy* distribution $P_{\rm me}$ is the most general, less structured model subject to the constraint (3.1) (and to no other constraint). The maximum entropy P results from the extremmum condition of the generalized entropy (its correct value can be shown to be a minimum):

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

Functional-derivating (3.2) with respect to $P(\mathbf{x})$ and equating to zero results in (the normalization will be ensured by a further Lagrange multiplier λ_0) (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 a Boltzmannian distribution in the canonical ensemble at temperature = 1, with effective Hamiltonian $\mathcal{H} = -\sum_k O_k$, but no assumption at all has been done on the Boltzmannian form, nor about thermal equilibrium, nor about the existence of an effective interaction (the Boltzmannian form is a consequence of the maximum entropy assumption –reflecting, rather, *absence* of hypothesis–).

The λ's in (3.3) are determined by optimizing (3.2) with respect to them. The result is (check!) again the maximum entropy condition, (3.1).

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!):

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

$$= -\langle \ln P_{\rm me} \rangle_{\rm e} = (1/M) \sum_{m=1}^M \ln P(\mathbf{x}^{(m)}) \tag{3.5}$$

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 (3.5, i.e., by maximising the *global, experimental likehood* according to the model). In this last scheme, the quantities $\langle O_k \rangle_{\rm e}$ are the *sufficient statistics* of the problem.

In its turn, this last equivalence can be viewed under an assumption of maximum likelihood, once one has assumed that the most probable distribution has the form (3.3). 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.

3.2 Examples of 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)):

$$P_{\rm me}(\mathbf{x}|J,\mathbf{h}) = \frac{1}{Z(J,\mathbf{h})} \exp\left[\sum_{i,j=1}^{n} J_{ij} x_i x_j + \sum_{i=1}^{n} h_i x_i\right] \tag{3.6}$$

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

$$\langle x_i x_j\rangle_P = \langle x_i x_j\rangle_{\rm e} \qquad \langle x_i\rangle_P = \langle x_i\rangle_{\rm e} \tag{3.7}$$

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.

Once one has solved the direct 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, as the microscopic information about the interaction; the possibility of recognizing novel thermalised configurations or creating new ones (learning); and estimating new observables, $\langle O\rangle_P$ if they do not have been measured. Indeed, a self-consistency test of the inference procedure is that of calculating different nontrivial, nonzero observables according to P, and comparing them with their experimental counterparts. Two-degree of freedom experimental correlations agree with those according to P by construction. But 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.

Exercise 5. *Does a model with pairwise interactions, as the one defined by (3.6) presents necessarily null higher-tan-2-point (connected) correlations?*

Example 2. Maximum likelihood inferring the coupling parameters of a statistical model. *There are actually (at least) two sources of errors in maximum entropy inference: the choice of the operators O and their number could lead to an underestimation of the data likelihood; even in the case that these are correct, one can infer poorly simply due to the ambiguity induced by the finiteness of the input data. Imagin a case in which the hypothesis are correct: as in the inverse problem of an interacting pairwise (for example a ferrogmagnetic) model: one would like to infer the coupling matrix J_{ij} of the model from a finite set of thermalized model configurations $\boldsymbol{\sigma}^{(k)}$. One should maximise the likelihood of the joint set of configurations, exactly as we did in the previous examples of inverse problem, differentiating (2.2,2.8) with respect to the model parameters and equating them to zero. This results in*

$$\frac{\partial}{\partial J_{ij}} \sum_k \ln P(\boldsymbol{\sigma}^{(k)}) = 0 \implies \langle\sigma_i\sigma_j\rangle_{\rm e} = \langle\sigma_i\sigma_j\rangle_P, \tag{3.8}$$

as we already know. 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:

$$\delta_{ij}(t) = -\eta \left.\frac{\partial}{\partial J_{ij}}\right|_{J(t)} \sum_k \ln P(\boldsymbol{\sigma}^{(k)}) = \eta \left[\langle\sigma_i\sigma_j\rangle_{\rm e} - \langle\sigma_i\sigma_j\rangle_{P(t)}\right] \tag{3.9}$$

we correct, hence, the J's with a quantity proportional to the violation of the equalities (3.7). 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 $\langle O_k\rangle_P$ one has to solve a direct problem for the current value of the parameters J.

In the following subsections one presents approximations overcoming the aforementioned problem, or situations in which the inverse problem can be solved exactly.

3.2.1 Inference of Ising degrees of freedom in the linear response approximation

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 expression (check it!):

$$\langle x_i \rangle = -\frac{\mathrm{d}F}{\mathrm{d}h_i} \tag{3.10}$$

$$\langle x_i x_j \rangle = \frac{1}{\beta}\frac{\mathrm{d}^2 F}{\mathrm{d}h_i \mathrm{d}h_j} + \langle x_i \rangle \langle x_j \rangle \tag{3.11}$$

where $F = -\ln Z/\beta$ is the free energy, and 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 1.11). This approximation, $F \simeq \tilde{F}$, will be called *linear response approximation* [Kappen and Rodríguez, 1998]. One can see that (check!):

$$\frac{\mathrm{d}\tilde{F}}{\mathrm{d}h_i} = \frac{\partial \tilde{F}}{\partial h_i} + \sum_{j=1}^{n} \frac{\mathrm{d}a_j}{\mathrm{d}h_i}\frac{\partial \tilde{F}}{\partial a_j} \tag{3.12}$$

$$\langle x_i \rangle \simeq \bar{x}_i \tag{3.13}$$

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:

$$\begin{aligned} \langle x_i x_j \rangle &= \langle x_i \rangle \langle x_j \rangle - \frac{1}{\beta} A_{ij}, \qquad A_{ij} \equiv \frac{\mathrm{d}\bar{x}_j}{\mathrm{d}h_i} \\ (A^{-1})_{ij} &= \delta_{i,j}\frac{\beta}{1-\bar{x}_i^2} - J_{ij} \end{aligned} \tag{3.14}$$

where the second line can be demonstrated (check!) derivating w.r.t h_i the equation for $\bar{x}_j$, 1.11.

Exercise 6. *Demonstrate the linear response equations, 3.14.*

3.2.2 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 ˜98% of the departures from statistical independence in parasol cells and ˜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].

3.2.3 Inference of a model with real degrees of freedom

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:

$$P(\mathbf{x}) = \frac{1}{Z}\exp\left[-\mathbf{x}^\dagger J\mathbf{x} - \mathbf{h}\mathbf{x}\right] \tag{3.15}$$

Z is the partition function, normalizing P. We would like to compute the non connected two-point correlator according to P, $C_{ij} = \langle x_i x_j\rangle_P$. To do so, we consider the Fourier transform of P:

$$\tilde{P}(\mathbf{q}) = \frac{1}{Z}\int [\mathrm{d}\mathbf{x}]\, e^{-\mathbf{x}^\dagger J\mathbf{x}-\mathbf{h}\mathbf{x}}\, e^{\imath\mathbf{q}^\dagger\mathbf{x}} \tag{3.16}$$

where $[\mathrm{d}\mathbf{x}] = \prod_{j=1}^d \mathrm{d}x_j$. Being J real and symmetric, it can be diagonalized, let us define its eigenvectors and eigenvalues as λ_j and $\mathbf{e}^{(j)}$, $j = 1,\ldots,d$. The coordinates in the base defined by the eigenvectors are $\mathbf{x}' = E\mathbf{x}$, where $E_{ij} = e_j^{(i)}$ and $EJE^\dagger = \mathrm{diag}(\boldsymbol{\lambda})$. We perform a change of variables in the integral (the Jacobian factor is one since the transformation defined by E is unitary):

$$\tilde{P}(\mathbf{q}) = \frac{1}{Z}\int [\mathrm{d}\mathbf{x}']\, e^{-\sum_j x_j'^2\lambda_j - \sum_j h_j' x_j' + \imath\sum_j q_j' x_j'} \tag{3.17}$$

where $q'_j = \sum_m e_m^{(j)} q_m$, or $\mathbf{q}' = E\mathbf{q}$, and idem with $\mathbf{h}' = E\mathbf{h}$. Gaussian-integrating results in:

$$\tilde{P}(\mathbf{q}) = \frac{1}{Z}\prod_{j|\lambda_j\neq 0}\left(\frac{\pi}{\lambda_j}\right)^{1/2} e^{\left[-q'^2_j + h'_j - 2\imath h'_j q'_j\right]/4\lambda_j} \tag{3.18}$$

(notice than the 0-eigenvalues give a unit contribution). Let us consider the logarithm of $\tilde{P}(\mathbf{q})$, it is (using the definition of $\mathbf{q}'$):

$$\ln\tilde{P}(\mathbf{q}) = -\ln Z + \frac{1}{2}\sum_{j|\lambda_j\neq 0}\ln\left[\pi/\lambda_j\right] - \sum_{m,n} q_m q_n (J^{-1})_{mn} + 2\imath\sum_m q_m \sum_n (J^{-1})_{mn} h_n + D \tag{3.19}$$

where $D = \sum_{m,n} h_m h_n (J^{-1})_{mn}$, and by J^{-1} we mean:

$$(J^{-1})_{mn} = \sum_{j|\lambda_j\neq 0} f_m^{(j)} f_n^{(j)} \lambda_j^{-1}; \tag{3.20}$$

the matrix J^{-1} coincides with the inverse matrix of J in the absence of null eigenvalues.

Let us now compare this expression for $\ln\tilde{P}(\mathbf{q})$ with its general expression (which can be found expanding the exponential in the definition of the Fourier transform):

$$\ln\tilde{P}(\mathbf{q}) = \imath\mathbf{q}^\dagger\langle\mathbf{x}\rangle_P - \frac{1}{2}\sum_{m,n} q_m q_n \langle x_m x_n\rangle_P + \mathcal{O}[q^3] \tag{3.21}$$

The comparison with (3.19) results in:

$$\langle x_i x_j\rangle_P = \frac{1}{2}(J^{-1})_{ij} \tag{3.22}$$

$$\langle x_i\rangle_P = 2\sum_j (J^{-1})_{ij} h_j \tag{3.23}$$

thus, the two-point correlation matrix is the inverse of the interaction matrix (as for the Ising model mean field approximation, c. f. 3.14).

3.2.4 Inferring the human aesthetic criteria in facial attractiveness

Consider a set of S human voters 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 $\mathbf{x}^{(k)}$, $k = 1,\ldots,S$ 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 $\{\mathbf{x}^{(k)}\}_{k=1}^S$ are real (positive and negative) D-dimensional vectors with zero mean on every component. One is, in principle, legitimated

to apply the pairwise maximum entropy method, whose valididy may be *a posteriori* assessed: one looks for the probability distribution on the faces $P(\mathbf{x})$, such that the experimental and theoretical two-facial distance correlations equal: $\langle x_i x_j \rangle_P = \langle x_i x_j \rangle_{\mathrm{e}}$. Using the results of the precedent subsection, one obtains that the effective coupling among couples of distances (proportions):

$$J = C_{\mathrm{e}}^{-1}$$
$$C_{\mathrm{e}ij} = \sum_{k=1}^{S} x_i^{(k)} x_j^{(k)},$$

where C_{e} is the experimental two-distance correlation among chosen faces. The J_{ij} matrix can be interpreted as *effective interactions* between distances in the subject's mind when he/she is committed in the beauty assessment, considered as a cognitive process.[4] This problem could provide an example of a system that is not properly inferrable by accounting pairwise correlations (hence, interactions) only. Indeed, the experimental three-distance interaction $\langle x_i x_j x_k \rangle_{\mathrm{e}}$ does not coincide with the equivalent quantity provided by the theoretical P in pairwise maximum entropy approximation (which, incidentally, vanishes). This suggests that one should improve the approximation, taking into account higher-order correlations: the resulting maximum entropy model effective Hamiltonian would exhibit $2+3$-distance interactions. In its turn, this may suggest [Ibanez-Berganza et al., 2017] that one evaluates facial properties which are more complicated than two-distance relations (i.e., proportions) in the facial attractiveness assessment.

Exercise 7. *Why should the 3-distance correlations be zero according to the pairwise maximum entropy approximation in this case (recall the Wick theorem)?*

3.2.5 Inference with $O(3)$ vectors in the spin-wave approximation

Consider a system of n agents collectively flying in three-dimensional space, whose velocity versors are $\{\mathbf{v}_i\}_{i=1}^n$ (the bold font denotes in this section spatial vectors in the three-dimensional unit sphere). The maximum entropy problem applied to the measurement of 2-point correlations $\langle \mathbf{v}_i \cdot \mathbf{v}_j \rangle$ leads to a model partition function:

$$Z(J) = \int [\mathrm{d}\mathbf{v}] \, \exp\left[\frac{1}{2}\sum_{ij} J_{ij} \mathbf{v}_i \cdot \mathbf{v}_j\right] \prod_{i=1}^{n} \delta(\mathbf{v}_i^2 - 1) \tag{3.24}$$

where $[\mathrm{d}\,\mathbf{v}] = \prod_{i=1}^n \mathrm{d}\mathbf{v}_i$. We define the total velocity $\mathbf{Y} = N\mathbf{y}$ of the flock, and the decomposition of each velocity along y: $\mathbf{v}_i = \mathbf{p}_i + \mathbf{y}\ell_i$. It is consequently $\sum_i \mathbf{p}_i = \mathbf{0}$. The partition function:

$$Z(J) = \int [\mathrm{d}\mathbf{p}] \int [\mathrm{d}\ell] \, \exp\left[\frac{1}{2}\sum_{ij} J_{ij} \left(\mathbf{p}_i \cdot \mathbf{p}_j + \ell_i \ell_j\right)\right] \delta\left(\sum_i \mathbf{p}_i\right) \prod_j \left[\delta(\ell_j^2 + \mathbf{p}_j^2 - 1)\right] \tag{3.25}$$

one approximates the argument in the second delta function as $\ell_i \simeq 1 - \mathbf{p}_i^2/2$, adding a Jacobian, $= \left[(1-\mathbf{p}_i^2)^{-1/2}\right]$ for each i, corresponding to the transformation from $\ell_i = (1-\mathbf{p}_i)^{1/2}$ to $\ell_i' = 1 - \mathbf{p}_i^2/2$. The longitudinal componets can be then integrated out. This results in (check!):

$$Z(J) = \int [\mathrm{d}\mathbf{p}] \, \exp\left[-\frac{1}{2}\sum_{ij} A_{ij} \mathbf{p}_i \cdot \mathbf{p}_j + \frac{1}{2} J_{ij}\right] \delta\left(\sum_i \mathbf{p}_i\right) \prod_j \left[\delta(1-\mathbf{p}_i^2)^{-1/2}\right] \tag{3.26}$$

where $A_{ij} = \sum_k J_{ik}\delta_{ij} - J_{ij}$ is the Laplacian matrix of the graph. When the system is very polarized (the perpendicular components are small), the last product of delta functions can be neglected, as argued in [Bialek et al., 2012]. Being real and symmetric, the Laplacian matrix can be diagonalized:

[4] Moreover, the principal components are axis in face space which are not interacting (the effective coupling among which is zero), hence they could represent (still an ongoing work) different, meaningful *independent* criteria [Ibanez-Berganza et al., 2017] of different experimental subjets.

$$\sum_j A_{ij} w_j^{(k)} = a_k w_i^{(k)} \tag{3.27}$$

There exist a null eigenvalue, $a_1 = 0$, corresponding to the constant eigenvector. The number of null eigenvalues corresponds to 1+the number of connected components of the graph (see for example [Anderson Jr and Morley, 1985]). In terms of them, the partition function:

$$Z(J) = e^{\sum_{ij} \frac{1}{2} J_{ij}} \int [\mathrm{d}\mathbf{p}'] \, \exp\left[-\frac{1}{2} \sum_{j=2}^{n} a_k (\mathbf{p}'_j)^2 \right] \delta\left(\mathbf{p}'_1\right) \tag{3.28}$$

where $\mathbf{p}'_i = \sum_j w_j^{(i)} \mathbf{p}_j$ (notice that the transformation $\mathbf{p}' \to \mathbf{p}$ exhibits unit determinant). Using Gaussian integration this leads (check!) to:

$$\ln Z(J) = -\sum_{j=2}^{n} \ln a_j + \frac{1}{2} \sum_{ij} J_{ij} + (n-1) \ln(2\pi)^{1/2} \tag{3.29}$$

and the 2-point correlator of the normal component of the velocity is (check!)[5]:

$$\langle \mathbf{p}_i \mathbf{p}_j \rangle_P = 2 \sum_{k=2}^{n} \frac{w_i^{(k)} w_j^{(k)}}{a_k} \tag{3.30}$$

(notice that the 2 factor comes from the two independent components of the $\mathbf{p}$'s. From this equation we learn that the correlation matrix is twice the inverse of matrix A[6].

3.2.6 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 $\langle \mathbf{p}_i \mathbf{p}_j \rangle_{\mathrm{e}}$ 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, $d = 1, 2, \cdots, n$:

$$C(\{\mathbf{v}\}, d) = \frac{1}{n} \sum_{i,j=1}^{n} \mathbf{v}_i \mathbf{v}_j \delta_{D_{ij},d} \tag{3.31}$$

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

$$-\sum_{d=1}^{N} J_d \; C(\{v\}, d) = -\sum_{i,j} J_{D_{ij}} \; \mathbf{v}_i \cdot \mathbf{v}_j \tag{3.32}$$

The analytical expression of the partition function:

$$\ln Z(J) = -\sum_{j=2}^{n} \ln a_j + n \sum_d J(d) \tag{3.33}$$

makes possible the maximization of the log likelihood:

$$\ln P = \langle \ln Z(J) + n \sum_d J(d) C(\{s\}, d) \rangle_{\mathrm{e}} \tag{3.34}$$

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. (3.33).

[5] Mind the identity $\int_{-\infty}^{\infty} x^{2n} e^{-\alpha x^2} = (\pi/\alpha)^{1/2} ((2n-1)!!)(2\alpha)^{-n}$.

[6] Mind that, in the basis of the $\mathbf{p}'$'s, the inverse of matrix A is, let us call it $(\tilde{A}^{-1})_{ij}$, is $= \delta_{ij} a_j^{-1}$. Hence $A^{-1} = U^\dagger \tilde{A}^{-1} U$ with $U_{ij} = w_i^{(j)}$.

3.2.7 Other variants of maximum entropy inference

We mention different studies of whith pairwise maximum entropy inference.

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

2. In [Cavagna et al., 2017], the maximum entropy with brute force exact calculation of the term $\ln Z(\boldsymbol{\lambda})$ in 3.5 has been used to infer causal relationships from experimental correlations between different patient symptoms and properties in medical diagnosis.

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

4. 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.

4 Neural networks: learning as inference

Consider, as before, a phase space, Σ, with M components (or "particles"), $\mathbf{x} = (x_1, \cdots, x_M)$, and consider a set of K empirical configurations $\mathbf{x}^{(k)} \in \Sigma$, $k = 1, \cdots, K$, 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:

$$\mathcal{L}(\mathbf{x}|W) = \exp[-\mathcal{H}[\mathbf{x}, W]]/Z(W) \tag{4.1}$$

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

$$\mathcal{H}(\mathbf{x}, W) = -\mathbf{x}^\dagger W \mathbf{x} \tag{4.2}$$

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!):

$$\frac{\partial \ln \mathcal{L}}{\partial W_{ij}} = K \left(x_i x_j - \langle x_i x_j \rangle_{\mathcal{L}} \right) \tag{4.3}$$

(as already seen in eq. 3.5). The W-gradient increases according to the *awake* (or experimental) 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) correlations (in the same way that the maximum entropy approach with pairwise correlations will efficiently capture the essential properties of systems that are essentially descrivable by two-point correlations). While in physics two-body correlations are often enough to efficiently describe the system, in a learning context they may be not enough[7]. To induce correlations between the M components of the model, N *hidden units* are introduced, which are additional variables of the model, to be marginalized when comparing the model with the experiment. The $N + M$ degrees of freedom of each configuration are now $\mathbf{x} = (\mathbf{v}, \mathbf{h})$, with $\mathbf{v} = (v_1, \cdots, v_M)$ and $\mathbf{h} = (h_1, \cdots, h_N)$:

$$\ln \mathcal{L}(\mathbf{v}|W) = \ln \sum_{\mathbf{h}} \exp[\mathbf{x}^\dagger W \mathbf{x}] - \ln Z(W) \tag{4.4}$$

$$Z(W) = \sum_{\mathbf{v},\mathbf{h}} \exp[\mathbf{x}^\dagger W \mathbf{x}] \tag{4.5}$$

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

$$\frac{\partial \ln \mathcal{L}(\mathbf{v}', W)}{\partial W_{ij}} = \sum_{\mathbf{h}} p_W(\mathbf{h}|\mathbf{v}')(x'_i x'_j) - \sum_{\mathbf{h},\mathbf{v}} p_W(\mathbf{v}, \mathbf{h})(x_i x_j) \tag{4.6}$$

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

4.1 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 $i, j \leq M$ and for all $i, j \geq M + 1$. Re-defining the out-diagonal matrix W as an N times M matrix, w, and considering external fields, the effective energy reads:

[7] an academic example is the *shifter ensemble*, an ensemble of strings of bits such that the second half is the first half shifted by a random number of bits to the left or to the right, with periodic boundary conditions. A Boltzmann machine with hidden units is not able to describe the ensemble with high likelihood

$$\mathcal{H}(\mathbf{v},\mathbf{h}|W,\mathbf{b},\mathbf{c}) = -\sum_{i=1}^{N}\sum_{j=1}^{M} w_{ij}h_i v_j - \sum_{i=1}^{N} h_i c_i - \sum_{j=1}^{M} v_j b_j \tag{4.7}$$

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!):

$$p(h_i = 1|\mathbf{v}) = \sigma\left(\sum_j w_{ij}v_j + c_i\right) \tag{4.8}$$

$$p(v_j = 1|\mathbf{h}) = \sigma\left(\sum_i w_{ij}h_i + b_j\right) \tag{4.9}$$

where $\sigma(y) = 1/(1+e^{-x})$. Thanks to this factorization, the first (awake) and second (sleep) terms in (4.6) are (check!):

$$\sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v})h_i v_j = p(h_i = 1|\mathbf{v})v_j \tag{4.10}$$

$$\sum_{\mathbf{v},\mathbf{h}} p(\mathbf{v},\mathbf{h})v_i h_j = \sum_{\mathbf{v}} p(\mathbf{v})p(h_i = 1|\mathbf{v})v_j \tag{4.11}$$

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

$$\frac{\partial \ln \mathcal{L}(\mathbf{x}', W)}{\partial w_{ij}} = p(h_i = 1|\mathbf{v}')v'_j - \sum_{\mathbf{v}} p(\mathbf{v})p(h_i = 1|\mathbf{v})v_j \tag{4.12}$$

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

$$\frac{\partial \ln \mathcal{L}_{\text{all}}(\mathcal{K}, W)}{\partial w_{ij}} = \frac{1}{K}\sum_{k=1}^{K} p(h_i = 1|\mathbf{v}^{(k)})v_j^{(k)} - \sum_{\mathbf{v}} p(\mathbf{v})p(h_i = 1|\mathbf{v})v_j \tag{4.13}$$

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

$$\frac{\partial \ln \mathcal{L}(\mathbf{x}', W)}{\partial b_j} = v'_j - \sum_{\mathbf{v}} p(\mathbf{v})v_j \tag{4.14}$$

$$\frac{\partial \ln \mathcal{L}(\mathbf{x}', W)}{\partial c_i} = p(h_1 = 1|\mathbf{v}') - \sum_{\mathbf{v}} p(\mathbf{v})p(h_1 = 1|\mathbf{v}) \tag{4.15}$$

In this way, one has reduced the complexity of the problem to the computation of only one ensemble average, that corresponding to the sleep term. This ensemble calculation (the second term in (4.13)) 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 $= \mathbf{v}^{(k)}$ for some $k = 1, \ldots, K$)
2. for $t = 0, \ldots, T$:
   - sample a vector $\mathbf{h}(t) \sim p(h_i|\mathbf{v}(t))$
   - sample a vector $\mathbf{v}(t+1) \sim p(v_i|\mathbf{h}(t))$

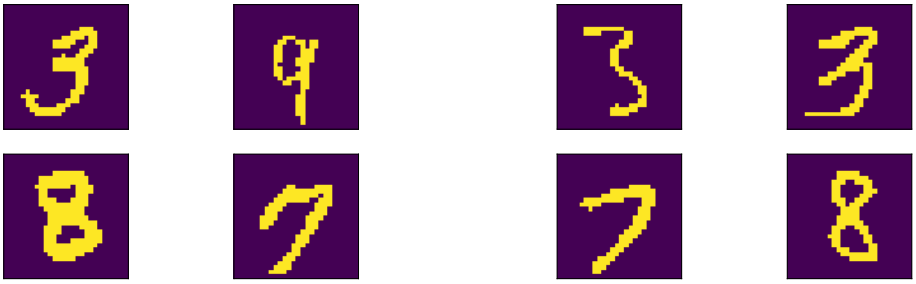

Figure 4.1: Four letters belonging to the training dataset of the MNIST ensemble (left), and four letters generated (with a *daydream* Gibbs algorithm) from the inferred likelihood $\mathcal{L}_{\rm all}$ (right). The learned letters seem handwritten.

3. approximate the sleep term in (4.13), $\sum_{\mathbf{v}} p(\mathbf{v}) p(h_i = 1|\mathbf{v}) v_j$, 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 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 (4.13), approximating the sleep term with the T-step contrastive divergence agorithm, schetched before, and call it δW_{ij}
   (b) as in (3.9),
   $$W_{ij}(\mathsf{r}+1) = W_{ij}(\mathsf{r}) - \eta\, \delta W_{ij} \tag{4.16}$$
   the η parameter is called the *learning rate*.

4.2 Two examples of unsupervised learning in RBM's

As an illustration of the learning process in a RBM, we present two applications of an RBM which (unsupervisedly) learns from a dataset according to the algorithm described above.

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. 4.1, 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 figure 4.2 we show the likelihood of 100 random MNIST test digits, while in fig. 4.1, right we show four random letters extracted from the learned $\mathcal{L}_{\rm all}$ (see figs. 4.2, right and 4.1, right).

In the first one, the RBM has learned the shifter ensemble (SE) (c. f. footnote 7). 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, or shifted by m positions, or shifted by $-m$ positions (i. e., by m positions at left), with periodic boundary conditions in the second half. For example, both 010100 and 010010 belong to the $(26, 1)$-SE. As we mentioned, such an ensamble is not learnable with maximum entropy. The RBM has learned an 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 (different from the K digits used in the training set), see fig. 4.2, left. The average likelihood of the dataset increases with the number of iterations, R; for a larger R the RBM would likely learn the esemble.

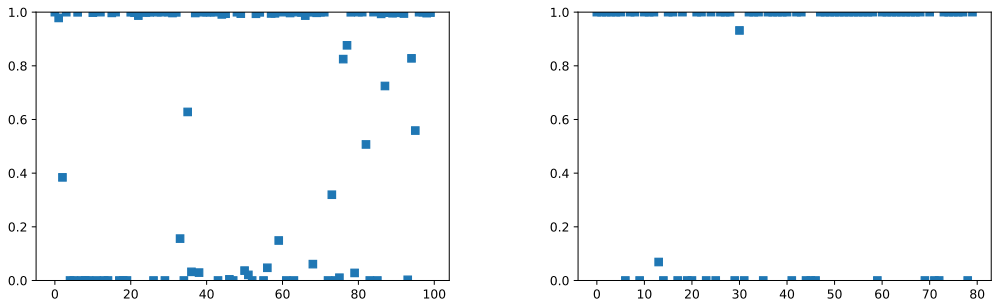

Figure 4.2: The learned likelihood of a test dataset in the cases of the MNIST (left) and $(26, 1)$-SE (right).

5 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* $p^{(0)}_{ij}$ (where $p^{(0)}$ is a stochastic irreducible matrix), and accepting it with probability a_{ij}. The transition matrix is hence $p_{ij} = a_{ij}p^{(0)}_{ij}$ for $i \neq j$ and $p_{ii} = p^{(0)}_{ii} + \sum_{j\neq i} p^{(0)}_{ij}(1 - a_{ij})$ (for the correct normalization it is necessary to keep refused configurations). Detailed balance is satisfied if

$$a_{ij} = F\left(\frac{\pi_j p^{(0)}_{ji}}{\pi_i p^{(0)}_{ij}}\right) \tag{5.1}$$

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

$$F(x) = \min\{x, 1\} \tag{5.2}$$

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:

$$a_{ij} = \min\left\{\frac{\pi_j}{\pi_i}, 1\right\} \tag{5.3}$$

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 $\boldsymbol{\sigma} = \otimes_{m=1}^{N}\sigma^{(m)}$, where $\sigma^{(m)}$ is the m-th particle state. Let $p^{(m)}$ be the transition matrix in which only particle m is updated:

$$p^{(m)}_{ij} > 0 \qquad \sigma^{(n)}_i = \sigma^{(n)}_j \qquad \forall n \neq m \tag{5.4}$$

$$p^{(m)}_{ij} = 0 \qquad \text{otherwise} \tag{5.5}$$

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).

5.1 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.

References

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

[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.

[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)*.

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

[Ibanez-Berganza et al., 2017] Ibanez-Berganza, M., Amico, A., and Loreto, V. (2017). *(work in progress)*.

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

[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.

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

[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.

[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*.

[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.

[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.