Inference, learning: common aspects

miguel.berganza@roma1.infn.it

Contents

Abstract

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

1 Inference: basic notions

1.1 Bayesian estimators

We remind Bayes theorem:

$$P(\theta|D) = \frac{\mathcal{L}(D|\theta)\pi(\theta)}{\mathcal{E}(D)} \tag{1.1}$$

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.

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 $\hat{\theta} = \arg\max_\theta \mathcal{L}(D|\theta)$.

[1] Part of the course *Sistemi Complessi*, Laurea Magistrale in Fisica, Università di Roma, "La Sapienza", anno accademico 2016-2017.

1.2 An example: 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:

$$\ln \mathcal{L}(D|\mu,\sigma) = -n \ln[(2\pi)^{1/2}\sigma] - [n(\mu-\bar{x})^2 + S]/(2\sigma^2) \tag{1.2}$$

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:

$$\mu^* = \bar{x} \tag{1.3}$$

$$\sigma^{*2} = n^{-1} S \tag{1.4}$$

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[2], the resulting ML estimator for σ results to be a biased estimator. The unbiased estimator is obtained by *marginalizing* the likelihood with respect to the mean:

$$\mathcal{L}(D|\sigma) = \int_{-\infty}^{\infty} \mathrm{d}\mu\, \mathcal{L}(D|\mu,\sigma)\pi(\mu) \tag{1.5}$$

$$\ln \mathcal{L}(D|\sigma) = -n \ln((2\pi)^{1/2}\sigma) - S/(2\sigma^2) + \ln((2\pi/n)^{1/2}\sigma/\sigma_\mu) \tag{1.6}$$

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!):

$$\sigma^{*2} = S/(n-1) \tag{1.7}$$

Exercise 1. *Obtain the marginal probability distribution for the average, $\mathcal{L}(D|\mu)$.*

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*:

$$\beta\tilde{F}(\boldsymbol{\theta},\beta,J) = \beta\langle E(\mathbf{x},J)\rangle_{Q(\mathbf{x};\boldsymbol{\theta})} - S[Q] = \tag{1.10}$$

$$\langle \ln \frac{Q(\mathbf{x};\boldsymbol{\theta})}{\mathcal{L}(\mathbf{x}|\beta,J)} \rangle_{Q(\mathbf{x};\boldsymbol{\theta})} - \ln Z(\beta,J) = \tag{1.11}$$

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

where $S[Q]$ is the entropy of the distribution Q, $\mathrm{KL}(A,B) = \langle \ln(A/B)\rangle_A$ is the relative entropy (or Kullback-Leibler divergence) between the distributions A and B, and $F = -\ln Z/\beta$ is the free

[2] 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.

energy. 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

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

where, J is a real symmetric matrix with zero diagonal, $\mathbf{h}$ is a real vector.

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 $\boldsymbol{\theta} = \mathbf{a} = \{a_i\}_{i=1}^n$:

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

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 $\bar{x}_i = 2q_i - 1$ 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 to the set of coupled equations:

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

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. 1.1 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.

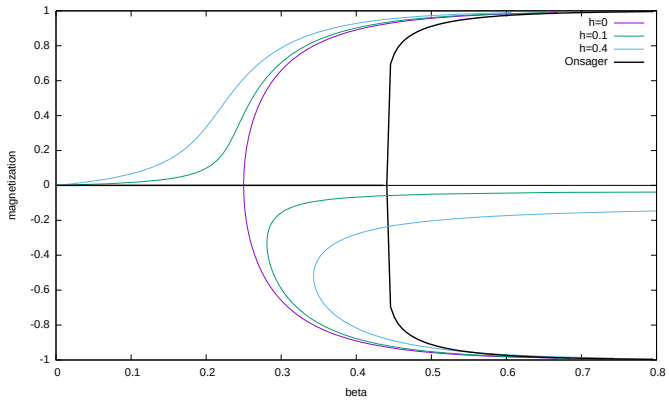

Figure 1.1: $\bar{m}(\beta)$.

Exercise 2. *Demonstrate Gibbs inequality, or* $\langle \ln(q/p) \rangle_q \geq 0$*, the equality being for* $q = p$ *only.*

1.4 Infering from two cues.

Suppose that the value of a quantity d (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 $\mathcal{L}(c|d)$. One could try to test experimentally whether a human subject is able to infer d from the value c, and whether she uses Bayes' rule. The test would require the knowledge of $\mathcal{L}(c|d)$.

Suppose, alternatively, that the subject is asked to infer the value of a quantity, d, from the values of *two cues* for this quantity, $c_{1,2}$. The posterior probability for d:

$$P(d|c_1, c_2) \propto \mathcal{L}(c_1, c_2|d)\pi(d) \tag{1.18}$$

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

$$P_i(d|c_i) = \mathcal{L}_i(c_i|d)\pi_i(d)/\mathcal{E}(c_i) \tag{1.19}$$

Now suppose that one can assume uniform priors and evidences. In this case, $P(d|c_1, c_2) \propto \prod_{i=1,2} P_i(d|c_i)$. Suppose, instead, that the subject has a prior tendency to attribute a value, $\pi_{1,2}(d) = p(d)$, only in the case on single-cue observations (i. e., π in (1.18) is still a uniform probability). In this case, it is:

$$P(d|c_1, c_2) \propto P_1(d|c_1)P_2(d|c_2)p(d)^{-2} \tag{1.20}$$

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

$$d^*(c_1, c_2) = \mu_1(c_1)w_1 + \mu_2(c_2)w_2 - 2\mu w \tag{1.21}$$

$$w_1 = \frac{\sigma_2^2\sigma^2}{A}, \qquad w_2 = \frac{\sigma_1^2\sigma^2}{A}, \qquad w = \frac{\sigma_1^2\sigma_2^2}{A} \tag{1.22}$$

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

Now, imagine that the parameters $\mu_{1,2}(c_{1,2})$ and $\sigma_{1,2}^2$, corresponding to the inference abilities of a given subject, can be estimated from a number of experimental estimations by the subject on the value of d given c_1 and c_2 in some set S. One then can present *both cues at the same time* to the subject, extracted from the set S, and see whether she infers d according to the optimal solution, (1.22). See ref. [Jacobs(1999)] for a pionneer implementation of such an experiment, and ref. [Knill and Pouget(2004)] for a review about bayesian brain functions.

1.5 Maximum entropy

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 $f_k : \Sigma \to \mathbb{R}$. The experimental averages are called $\langle f_k \rangle_{\mathrm{e}} = \frac{1}{M}\sum_{i=1}^M f_k^{(i)}$, where $f_k^{(i)}$ is the i-th experimental result, $i = 1, \cdots, M$, 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:

$$\langle f_k \rangle_P = \langle f_k \rangle_{\mathrm{e}}. \tag{1.23}$$

In other words the *maximum entropy* distribution P_{me} is the most general, less structured model subject to the constraint (1.23) (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):

$$\mathcal{S}[P] \equiv S[P] + \sum_{k=1}^{K} \lambda_k (\langle f_k \rangle_{\mathrm{e}} - \langle f_k \rangle_P) \tag{1.24}$$

Functional-derivating (1.24) 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!):

$$P_{\mathrm{me}}(\mathbf{x}) = \frac{1}{Z(\boldsymbol{\lambda})} \exp\left[\sum_{k=1}^{K} \lambda_k f_k(\mathbf{x})\right] \tag{1.25}$$

$Z(\boldsymbol{\lambda}) = \exp(-\lambda_0 - 1)$ being the normalizing constant. The λ's are determined by optimizing with respect to them. The result is (check!) again the maximum entropy condition, (1.23). Notice that, alternatively, the minimization of the generalized entropy is equivalent to the maximization of the experimental average of the likelihood (from eq. 1.24):

$$\mathcal{S}[P] = -\langle \ln P \rangle_{\mathrm{e}} = \ln Z(\boldsymbol{\lambda}) - \sum_{k=1}^{K} \lambda_k \langle f_k \rangle_{\mathrm{e}} \tag{1.26}$$

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 f of sec. 1.5 are averages and correlators (i. e., 1- and 2-point operators respectively): x_i, $x_i x_j$. The probability distribution on Σ results to be (c. f. (1.25):

$$P_{\mathrm{me}}(\mathbf{x}|J, \mathbf{h}) = \frac{1}{Z(\boldsymbol{\lambda})} \exp\left[\sum_{i,j=1}^{n} J_{ij} x_i x_j + \sum_{i=1}^{n} h_i x_i\right] \tag{2.1}$$

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

$$\langle x_i x_j \rangle_P = \langle x_i x_j \rangle_{\mathrm{e}} \qquad \langle x_i \rangle_P = \langle x_i \rangle_{\mathrm{e}} \tag{2.2}$$

where $\langle \rangle_{\mathrm{e}} \cdot$ refers to the experimental average. The problem is, in general, hard, since the evaluation of $\langle \cdot \rangle_P$ requires the solution of the statistical problem for the given value of the parameters $J, \mathbf{h}$. In the following subsections one presents approximations overcoming the problem.

2.1 Inference 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!):

$$\langle x_i \rangle = -\frac{\mathrm{d}F}{\mathrm{d}h_i} \tag{2.3}$$

$$\langle x_i x_j \rangle = \frac{1}{\beta} \frac{\mathrm{d}^2 F}{\mathrm{d}h_i \mathrm{d}h_j} + \langle x_i \rangle \langle x_j \rangle \tag{2.4}$$

where $F = -\ln Z/\beta$ is the free energy, and approximate F by the minimum of $\tilde{F}$ in sec. 1.5, 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!):

$$\frac{\mathrm{d}\tilde{F}}{\mathrm{d}h_i} = \frac{\partial \tilde{F}}{\partial h_i} + \sum_{j=1}^{n} \frac{\mathrm{d}a_j}{\mathrm{d}h_i} \frac{\partial \tilde{F}}{\partial a_j} \tag{2.5}$$

$$\langle x_i \rangle \simeq \bar{x}_i \tag{2.6}$$

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. 1.5. Oppositely, the correlations:

$$\begin{aligned} \langle x_i x_j \rangle = \langle x_i \rangle \langle x_j \rangle - \frac{1}{\beta} A_{ij}, \qquad A_{ij} \equiv \frac{\mathrm{d}\bar{x}_j}{\mathrm{d}h_i} \\ (A^{-1})_{ij} = \delta_{i,j} \frac{\beta}{1-\bar{x}_i^2} - J_{ij} \end{aligned} \tag{2.7}$$

where the second line can be demonstrated (check!) derivating w.r.t h_i the equation for $\bar{x}_j$, 1.16.

Exercise 3. *Demonstrate the linear response equations, 2.7.*

See much more on the mean field approximation applied to pairwise probability infering in [Roudi et al.(2009)Roudi, Aurell, and Hertz].

2.2 Inference with $O(3)$ vectors and 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 $\langle \mathbf{v_i v_j} \rangle$ leads to a model partition function:

$$Z(J) = \int [\mathrm{d}\mathbf{v}] \, \exp \left[\frac{1}{2} \sum_{ij} J_{ij} \mathbf{v}_i \cdot \mathbf{v}_j \right] \prod_{i=1}^n \delta(\mathbf{v}_i^2 - 1) \tag{2.8}$$

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:

$$Z(J) = \int [\mathrm{d}\mathbf{p}] \int [\mathrm{d}\ell] \, \exp \left[\frac{1}{2} \sum_{ij} J_{ij} \left(\mathbf{p}_i \cdot \mathbf{p}_j + \ell_i \ell_j \right) \right] \delta \left(\sum_i \mathbf{p}_i \right) \prod_j \left[\delta(\ell_j^2 + \mathbf{p}_j^2 - 1) \right] \tag{2.9}$$

one approximates the argument in the second delta function as $\ell_i \simeq 1 - \mathbf{p}_i^2/2$ and tranforms consequently the delta function over the ℓ's. This results in (the last product of delta functions can be neglected, see [Bialek et al.(2012)Bialek, Cavagna, Giardina, Mora, Silvestri, Viale, and Walczak]):

$$Z(J) = \int [\mathrm{d}\mathbf{p}] \int [\mathrm{d}v^\ell] \, \exp \left[-\frac{1}{2} \sum_{ij} A_{ij} \mathbf{p}_i \cdot \mathbf{p}_j + \frac{1}{2} J_{ij} \right] \delta \left(\sum_i \mathbf{p}_i \right) \prod_j \left[(1 - \mathbf{p}_i^2)^{-1/2} \right] \tag{2.10}$$

where $A_{ij} = \sum_k J_{ik} \delta_{ij} - J_{ij}$ is the Laplacian matrix of the graph. Being real and symmetric, it can be diagonalized:

$$\sum_j A_{ij} w_j^{(k)} = a_k w_i^{(k)} \tag{2.11}$$

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:

$$Z(J) = e^{\sum_{ij} \frac{1}{2} J_{ij}} \int [\mathrm{d}\mathbf{p}'] \, \exp \left[-\frac{1}{2} \sum_{j=2}^n a_k (\mathbf{p}'_j)^2 \right] \delta \left(\mathbf{p}'_1 \right) \tag{2.12}$$

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:

$$\ln Z(J) = -\sum_{j=2}^{n} \ln a_j + \frac{1}{2}\sum_{ij} J_{ij} + (n-1)\ln(2\pi)^{1/2} \tag{2.13}$$

and the 2-point correlator of the normal component of the velocity is (check!)[3]:

$$\langle \mathbf{p}_i \mathbf{p}_j \rangle_P = 2\sum_{k=2}^{n} \frac{w_i^{(k)} w_j^{(k)}}{a_k} \tag{2.14}$$

(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[4].

In reference [Cavagna et al.(2015)Cavagna, Del Castello, Dey, Giardina, Melillo, Parisi, and Viale] (see also [Bialek et al.(2012)Bialek, Cavagna, Giardina, Mora, Silvestri, Viale, and Walczak]) 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. 1.5, it is used $C(\{\mathbf{v}\}, d)$, the velocity correlation between between two birds at different topological distancies, $d = 1, 2, \cdots, n$:

$$C(\{\mathbf{v}\}, d) = \frac{1}{n}\sum_{i,j=1}^{n} \mathbf{v}_i \mathbf{v}_j \delta_{D_{ij},d} \tag{2.15}$$

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!):

$$-\sum_{d=1}^{N} J_d \; C(\{v\}, d) = -\sum_{i,j} J_{D_{ij}} \; \mathbf{v}_i \cdot \mathbf{v}_j \tag{2.16}$$

The analytical expression of the partition function:

$$\ln Z(J) = -\sum_{j=2}^{n} \ln a_j + n\sum_{d} J(d) \tag{2.17}$$

makes possible the maximization of the log likelihood:

$$\ln P = \langle \ln Z(J) + n\sum_{d} J(d) C(\{s\}, d) \rangle_{\mathrm{e}} \tag{2.18}$$

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. (2.17).

2.3 Other variants of maximum entropy inference

We mention two different studies of inference which pairwise correlations. In [Sakellariou et al.(2016)Sakellariou, Tria, Loreto, and Pachet], an improvement of mean field maximum entropy, called pseudo-likelihood inference, is used to capture statistics of melodies in music. In [Cavagna et al.(2017)Cavagna, Giradina, Skert, and Viale], the maximum entropy with brute force exact calculation of the term $\ln Z(\boldsymbol{\lambda})$ in 1.26 has been used to infer causal relationships from experimental correlations between different patient symptoms and properties in medical diagnosis.

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

[4] 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 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:

$$\mathcal{L}(\mathbf{x}|W) = \exp[-\mathcal{H}[\mathbf{x}, W]]/Z(W) \tag{3.1}$$

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

$$\mathcal{H}(\mathbf{x}, W) = -\mathbf{x}^\dagger W \mathbf{x} \tag{3.2}$$

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!):

$$\frac{\partial \ln \mathcal{L}}{\partial W_{ij}} = K \left(\langle x_i x_j \rangle_{\mathrm{e}} - \langle x_i x_j \rangle_{\mathcal{L}} \right) \tag{3.3}$$

(note that the argument leading to this equation is the specular of the argument leading to eq. 1.26). 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. 1.5.

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[5]. 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 $\mathbf{h} = (h_1, \cdots, h_N)$:

$$\ln \mathcal{L}(\mathbf{v}|W) = \ln \sum_{\mathbf{h}} \exp[\mathbf{x}^\dagger W \mathbf{x}] - \ln Z(W) \tag{3.4}$$

$$Z(W) = \sum_{\mathbf{v},\mathbf{h}} \exp[\mathbf{x}^\dagger W \mathbf{x}] \tag{3.5}$$

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

$$\frac{\partial \ln \mathcal{L}(\mathbf{x}, W)}{\partial W_{ij}} = \sum_{\mathbf{h}} p_W(\mathbf{h}|\mathbf{v})(x_i x_j) - \sum_{\mathbf{h},\mathbf{v}} p_W(\mathbf{v}, \mathbf{h})(x_i x_j) \tag{3.6}$$

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

3.1 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:

[5] 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

$$\mathcal{H}(\mathbf{v},\mathbf{h},W,\mathbf{b},\mathbf{c}) = -\sum_{i=1}^{N}\sum_{j=1}^{M} w_{ij}h_i v_j - \sum_{i=1}^{N} h_i c_i - \sum_{j=1}^{M} v_j b_j \tag{3.7}$$

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!):

$$p(h_i = 1|\mathbf{v}) = \sigma\left(\sum_j w_{ij}v_j + c_i\right) \tag{3.8}$$

$$p(v_j = 1|\mathbf{h}) = \sigma\left(\sum_i w_{ij}h_i + b_j\right) \tag{3.9}$$

where $\sigma(y) = 1/(1+e^{-x})$. Thanks to this factorization, the first (awake) and second (sleep) terms in (3.6) are (check!):

$$\sum_{\mathbf{h}} p(\mathbf{h}|\mathbf{v})h_i v_j = p(h_i = 1|\mathbf{v})v_j \tag{3.10}$$

$$-\sum_{\mathbf{v},\mathbf{h}} p(\mathbf{v},\mathbf{h})v_i h_j = \sum_{\mathbf{v}} p(\mathbf{v})p(h_i = 1|\mathbf{v})v_j \tag{3.11}$$

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

$$\frac{\partial \ln \mathcal{L}(\mathbf{x},W)}{\partial w_{ij}} = p(h_i = 1|\mathbf{v})v_j - \sum_{\mathbf{v}} p(\mathbf{v})p(h_i = 1|\mathbf{v})v_j \tag{3.12}$$

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

$$\frac{\partial \ln \mathcal{L}(\mathcal{K},W)}{\partial w_{ij}} = \frac{1}{K}\sum_{k=1}^{K} p(h_i = 1|\mathbf{v}^{(k)})v_j^{(k)} - \sum_{\mathbf{v}} p(\mathbf{v})p(h_i = 1|\mathbf{v})v_j \tag{3.13}$$

the log-likelihood derivative w.r.t. the fields are (check!):

$$\frac{\partial \ln \mathcal{L}(\mathbf{x},W)}{\partial b_j} = v_j - \sum_{\mathbf{v}} p(\mathbf{v})v_j \tag{3.14}$$

$$\frac{\partial \ln \mathcal{L}(\mathbf{x},W)}{\partial c_i} = p(h_1 = 1|\mathbf{v}) - \sum_{\mathbf{v}} p(\mathbf{v})p(h_1 = 1|\mathbf{v}) \tag{3.15}$$

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 can be performed with the help of a Gibbs Monte-Carlo algorithm, as illustrated in the RBM example.

References

[Anderson Jr and Morley(1985)] Anderson Jr, W. N. and T. D. Morley, 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, doi:10.1561/2200000006, URL http://dx.doi.org/10.1561/2200000006.

[Bialek et al.(2012)Bialek, Cavagna, Giardina, Mora, Silvestri, Viale, and Walczak] Bialek, W., A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, 2012: Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences*, **109 (13)**, 4786–4791, doi:10.1073/pnas.1118633109, URL http://www.pnas.org/content/109/13/4786.abstract, http://www.pnas.org/content/109/13/4786.full.pdf.

[Cavagna et al.(2015)Cavagna, Del Castello, Dey, Giardina, Melillo, Parisi, and Viale] Cavagna, A., L. Del Castello, S. Dey, I. Giardina, S. Melillo, L. Parisi, and M. Viale, 2015: Short-range interactions versus long-range correlations in bird flocks. *Phys. Rev. E*, **92**, 012 705, doi:10.1103/PhysRevE.92.012705, URL http://link.aps.org/doi/10.1103/PhysRevE.92.012705.

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

[Fischer and Igel(2012)] Fischer, A. and C. Igel, 2012: *An Introduction to Restricted Boltzmann Machines*, 14–36. Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-642-33275-3_2, URL http://dx.doi.org/10.1007/978-3-642-33275-3_2.

[Jacobs(1999)] Jacobs, R. A., 1999: Optimal integration of texture and motion cues to depth. *Vision Research*, **39 (21)**, 3621 – 3629, doi:http://dx.doi.org/10.1016/S0042-6989(99)00088-7, URL http://www.sciencedirect.com/science/article/pii/S0042698999000887.

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

[Knill and Pouget(2004)] Knill, D. C. and A. Pouget, 2004: The bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, **27 (12)**, 712 – 719, doi:http://dx.doi.org/10.1016/j.tins.2004.10.007, URL http://www.sciencedirect.com/science/article/pii/S0166223604003352.

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

[Roudi et al.(2009)Roudi, Aurell, and Hertz] Roudi, Y., E. Aurell, and J. A. Hertz, 2009: Statistical physics of pairwise probability models. *Frontiers in computational neuroscience*, **3**, 22, doi:10.3389/neuro.10.022.2009, URL http://europepmc.org/articles/PMC2783442.

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