EUROPHYSICS LETTERS

Europhys. Lett., (), pp. ()

Bayes-optimal performance in a discrete space

M. Copelli[1], C. Van den Broeck[1] and M. Opper[2]

[1] *Limburgs Universitair Centrum, B-3590 Diepenbeek, Belgium*
[2] *Neural Computing Research Group, Aston University, B4 7ET Birmingham, UK*

(received ; accepted)

PACS. 64.60Cn – Order disorder transformations; statistical mechanics of model systems.
PACS. 87.10+e – General theory and mathematical aspects.
PACS. 02.50−r – Probability theory, stochastic processes, and statistics.

Abstract. – We study a simple model of unsupervised learning where the single symmetry breaking vector has binary components ±1. We calculate exactly the Bayes-optimal performance of an estimator which is required to lie in the same discrete space. We also show that, except for very special cases, such an estimator cannot be obtained by minimization of a class of variationally optimal potentials.

Statistical mechanics techniques have been used with success to study and understand key properties of inferential learning [1, 2]. This approach provides explicit and detailed results that are in many ways complementary to the more general, but also more blend results obtained by statistics. The case of non-smooth problems, in which the parameters that have to be estimated take discrete values, is of particular interest. On the one hand, many of the results from statistics can no longer be applied, while on the other hand, the estimation of these parameters is often a computationally hard problem. In this paper, we present a detailed analysis of a simple model of unsupervised learning [3, 4, 5, 6, 7, 8], involving a single symmetry breaking vector with binary components ±1 and highlight the differences with the case of smooth components. In particular we compare the results from Gibbs learning and Bayes learning with the ones for the best binary vector and a vector which minimizes a variationally optimal potential.

The problem is as follows: a set of p N-dimensional patterns $\{\boldsymbol{\xi}^\mu,\ \mu = 1, ..., p\}$, are sampled independently from a distribution $P(\boldsymbol{\xi}^\mu|\mathbf{B}) \sim \delta(\boldsymbol{\xi}^\mu \cdot \boldsymbol{\xi}^\mu - N)\ \exp\left[-U\left(\mathbf{B}\cdot\boldsymbol{\xi}^\mu/\sqrt{N}\right)\right]$ with a single symmetry breaking direction $\mathbf{B}$. The function U modulates the distribution of the patterns along $\mathbf{B}$. We will focus on the properties in the thermodynamic limit $N \to \infty$, $p \to \infty$ with $\alpha = p/N$ finite. One then finds that the normalized projection $t \equiv \mathbf{B}\cdot\boldsymbol{\xi}/\sqrt{N}$ is distributed according to ($\mathcal{N}$ being a normalization constant)

$$P^*(t) = \frac{\mathcal{N}}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2} - U(t)\right\} , \tag{1}$$

Typeset using EURO-TEX

while projections on any direction orthogonal to **B** are normal. The case of a so-called spherical prior, in which **B** is chosen at random on the sphere with radius $\sqrt{N}$, was discussed in [6, 7, 8]. As announced earlier, we focus here on the more complicated situation in which the components of **B** take binary values ±1. The *prior distribution* is now given by:

$$P(\mathbf{B}) \equiv P_b(\mathbf{B}) = \prod_{j=1}^{N} \left[\frac{1}{2}\delta(B_j - 1) + \frac{1}{2}\delta(B_j + 1) \right] . \tag{2}$$

The goal of unsupervised learning is to give an estimate **J** of **B**. One way to do so is to sample **J** from a Boltzmann distribution with Hamiltonian $\mathcal{H}(\mathbf{J}) = \sum_{\mu=1}^{p} V(\lambda^\mu)$, with $\lambda^\mu \equiv \mathbf{J} \cdot \boldsymbol{\xi}^\mu / \sqrt{N}$, at temperature $T = \beta^{-1}$, for an appropriate choice of the ad-hoc potential V [9]. The properties of such a **J**-vector can be extracted from the partition function :

$$Z = \int d\mathbf{J}\, P_b(\mathbf{J})\, e^{-\beta \mathcal{H}(\mathbf{J})} . \tag{3}$$

The latter is a fluctuating quantity due to the random choice of the patterns, but the free energy per component $f = -\beta^{-1} \ln Z/N$ is expected to be self-averaging in the thermodynamic limit and can therefore be calculated by averaging over the pattern distribution with the aid of the replica trick [14]. Assuming replica symmetry (RS), one finds

$$\begin{aligned} f \quad &= \frac{1}{\beta} \operatorname*{Extr}_{R,q,\hat{R},\hat{q}} \left\{ \frac{1}{2}(1-q)\hat{q} + \hat{R}R - \int \mathcal{D}z \, \ln\cosh\left(z\sqrt{\hat{q}} + \hat{R} \right) \right. \\ &\left. -\alpha \int \mathcal{D}^*t \int \mathcal{D}t' \, \ln \int \frac{d\lambda}{\sqrt{2\pi(1-q)}} \exp\left(-\beta V(\lambda) - \frac{(\lambda - t'\sqrt{q - R^2} - tR)^2}{2(1-q)} \right) \right\} . \end{aligned} \tag{4}$$

where $\mathcal{D}^*t = dt\, P^*(t)$ and $\mathcal{D}t' = dt'\,(2\pi)^{-1/2} \exp(-t'^2/2)$. The extremum operator renders saddle point equations which determine the self-averaging value of the order parameters. As usual q can be interpreted as the typical mutual overlap between two samplers **J** and **J**′, $q = \mathbf{J} \cdot \mathbf{J}'/N$, while R measures the proximity between the estimate **J** and the "true" direction **B**, $R = \mathbf{J} \cdot \mathbf{B}/N$. For even functions U, there is no distinction between **B** and −**B**, and a symmetry $R \to -R$ arises. In the following, only $R \geq 0$ will be considered.

As a first application of eq. 4, we turn to Gibbs learning [15, 10, 5]. It corresponds to sampling from the posterior distribution and is realized by taking $\beta = 1$ and $V = U$ in eq. 4 [6]. In agreement with the fact that one cannot make a statistical distinction between **B** and its Gibbsian estimate **J**, one finds that the order parameters satisfy $q_G = R_G$ and $\hat{q}_G = \hat{R}_G$, where the subscript G refers to Gibbs learning. This observation allows to simplify the saddle point equations further, and the Gibbs overlap is found to obey the following equation:

$$R_G = F_B^2 \left(\mathcal{F}\left(\sqrt{R_G} \right) \right) \tag{5}$$

with

$$F_B(x) = \sqrt{\int \mathcal{D}z \, \tanh\left(zx + x^2 \right)} \quad \text{and} \quad \mathcal{F}(R) = \sqrt{\alpha \int \mathcal{D}t \, \frac{Y^2(t;R)}{X(t;R)}} \tag{6}$$

and

$$X(t;R) = \int \mathcal{D}t' \, \mathcal{N} e^{-U(Rt+\sqrt{1-R^2}t')} \quad Y(t;R) = \frac{1}{R}\frac{\partial}{\partial t} X(t;R) \; . \tag{7}$$

Note that F_B comes from the entropic term of the free energy and does not depend on U, as opposed to $\mathcal{F}$, X and Y.

For R_G small, one obtains from eqs. 5-7, upon assuming a smooth behaviour as a function of α, that ($\int \mathcal{D}^* t \; f(t) = \langle f(t) \rangle_*$):

$$\langle t \rangle_* \neq 0 \quad \Rightarrow \quad R_G \simeq \alpha \; \langle t \rangle_*^2 \tag{8}$$

$$\langle t \rangle_* = 0 \quad \Rightarrow \quad R_G \begin{cases} = 0, & \alpha \leq \alpha_G \\ \simeq C(\alpha - \alpha_G), & \alpha \geq \alpha_G \end{cases} \tag{9}$$

with critical load $\alpha_G = \left(1 - \langle t^2 \rangle_*\right)^{-2}$. These results are identical to those for a spherical prior [8]. In particular, one observes the appearance of *retarded learning* when the distribution has a zero mean along the symmetry breaking axis. In the regime $R_G \to 1$ on the other hand, one finds an exponential approach :

$$1 - R_G(\alpha) \overset{\alpha \to \infty}{\simeq} \sqrt{\frac{\pi}{2\alpha \langle (U')^2 \rangle_*}} \exp\left(\frac{-\alpha \langle (U')^2 \rangle_*}{2} \right) \; , \tag{10}$$

where $U' \equiv dU(t)/dt$. This is now different from the case of a spherical prior, where the approach is following an inverse power law $1 - R_G \sim \alpha^{-1}$ [8]. The difference becomes even more pronounced when U has singular derivatives, as is typically the case for supervised problems. Then one finds that $R_G = 1$ is attained at a *finite* value of α while $1 - R_G \sim \alpha^{-2}$ for a spherical prior, see [15] and [6] for an explicit example.

Apart from its intrinsic interest, Gibbs learning also provides an upper bound for the performance of any cost function via its relation with the Bayes-optimal overlap $R_B = \sqrt{R_G}$, see [16, 10, 5]. This performance is attained by the center of mass $\mathbf{J}_B$ of the Gibbs ensemble, which maximizes the overlap R averaged over the posterior distribution of $\mathbf{B}$, as can be seen from a simple reasoning [10, 5]. In order to exclude the trivial result $\mathbf{J}_B = 0$ (which would follow from the symmetry of the prior), we will implicitely assume an infinitesimally small symmetry breaking field in the Gibbs distribution.

Using the self-averaging of the mutual overlap, with $q_G = R_G$, the explicit form of $\mathbf{J}_B$ is found to be $\mathbf{J}_B = R_G^{-1/2} Z^{-1} \int d\mathbf{J} \, P_b(\mathbf{J}) \, \mathbf{J} \, \exp\{-\sum_\mu U(\lambda^\mu)\}$. In general, the components of this center of mass are continuous, while our prime interest here is in the optimal performance attainable by an Ising vector. The latter vector, which we will denote by $\mathbf{J}_{bb}$ (for best binary), can fortunately be easily obtained [1]: it is the clipped version of the center of mass $\mathbf{J}_B$, with components $(\mathbf{J}_{bb})_j = \mathrm{sign}\left((\mathbf{J}_B)_j\right)$.

To evaluate the performance of $\mathbf{J}_{bb}$, we recall the following general result for the overlap $\tilde{R} = \tilde{\mathbf{J}} \cdot \mathbf{B}/N$ of a vector $\tilde{\mathbf{J}}$ with transformed components $\tilde{J}_i = \sqrt{N} g(J_i)/\sqrt{\sum_i g^2(J_i)}$ (with g uneven and $\mathbf{B}$ Ising) as a function of the overlap R of $\mathbf{J}$ with $\mathbf{B}$ [17]:

$$\tilde{R} = \frac{\int P(x) \, g(x) \, dx}{\left[\int P(x) \, g^2(x) \, dx\right]^{1/2}} \; , \tag{11}$$

where $P(x)$ is the probability density for $x \equiv J_1 B_1$, which is independent of the index due to the permutation symmetry among the axes. If $\mathbf{J}$ is sampled from a spherical distribution (with $\mathbf{B}$ Ising), then $P(x)$ is found to be a Gaussian [17] with mean R and variance $1-R^2$. In order to obtain $P(x)$ corresponding to the center of mass $\mathbf{J}_B$, we evaluate the quenched moments of $y = x\sqrt{R_G}$:

$$\langle y^m \rangle = \left\langle \left(Z^{-1} \int d\mathbf{J}\, P_b(\mathbf{J})\, e^{-\sum_\mu U(\lambda^\mu)} J_1 B_1 \right)^m \right\rangle . \tag{12}$$

The average over the pattern set can again be performed by the replica trick with the following replica symmetric result:

$$\langle y^m \rangle = \int \mathcal{D}z \left[\tanh\left(z\sqrt{\hat{R}_G} + \hat{R}_G \right) \right]^m , \tag{13}$$

where $\hat{R}_G$, which is determined by the saddle point equations of Gibbs learning, cf. eq. 5, is found to be $\hat{R}_G = \mathcal{F}^2(\sqrt{R_G})$. Recognizing eq. 13 as a transformation of variables $y = \tanh\left(z\sqrt{\hat{R}_G} + \hat{R}_G \right)$, with z normally distributed, one concludes [18] :

$$P(x) = \frac{\sqrt{R_G}}{\sqrt{2\pi\hat{R}_G}(1 - R_G\, x^2)} \exp\left\{ \frac{-1}{2\hat{R}_G} \left[\frac{1}{2} \ln\left(\frac{1+\sqrt{R_G}\, x}{1-\sqrt{R_G}\, x} \right) - \hat{R}_G \right]^2 \right\} . \tag{14}$$

By applying eq. 11, for $g(x) = \mathrm{sign}(x)$, with $P(x)$ given by eq. 14, one finally obtains the following overlap $R_{bb} \equiv \mathbf{J}_{bb} \cdot \mathbf{B}/N$ of the best binary vector :

$$R_{bb} = 1 - 2H\left(F_B^{-1}(R_B) \right) = 1 - 2H\left(\mathcal{F}\left(R_B\right) \right) . \tag{15}$$

Eq. 15 is a central result of this paper, providing an upper bound for the performance of any binary vector. The asymptotics of R_{bb} can be obtained from those of $R_G = R_B^2$, yielding

$$R_{bb} \overset{R_G \to 0}{\simeq} \sqrt{\frac{2R_G}{\pi}} \tag{16}$$

in the poor performance regime, and an exponential behavior in the limit of $R_G \to 1$:

$$1 - R_{bb} \simeq \frac{2}{\pi}(1 - R_G) . \tag{17}$$

We note that another quantity of interest, the mutual overlap $\Gamma \equiv \mathbf{J}_B \cdot \mathbf{J}_{bb}/N$ between center of mass and best binary, can also be evaluated quite easily, leading to the simple result $\Gamma = R_{bb}/R_B$. In the limit $R_G \to 0$ one recovers $\Gamma \to \sqrt{2/\pi}$, which is the result for the overlap between a vector sampled at random from the N-sphere and its clipped counterpart. Γ, R_B and R_{bb} are plotted as functions of R_G in fig. 1.

Fig. 1. – Γ, R_B and R_{bb} parametrized by R_G, according to eqs. 5 and 15.

We finally turn to the problem of a variationally optimized potential. In the case of a spherical prior, it was shown that the Bayes-optimal performance can indeed be attained by a vector that minimizes this potential [11, 8, 12, 13]. We now address the question of whether the same procedure is successful in discrete space, a problem which has been also studied in [19] for the supervised scenario. Since $\mathbf{J}_{bb}$ is a unique optimal binary vector, the desired potential would have to satisfy both $R = R_{bb}$ and $q = 1$ (otherwise one would be able to construct the center of mass of this ensemble with yet a larger overlap, violating the bound). Proceeding again from the free energy eq. 4 for a general potential V, taking the limits $q \to 1$, $\beta \to \infty$ with finite $c \equiv \beta(1-q)$, and rescaling the conjugate parameters $\hat{c} \equiv \hat{q}/\beta^2$, $\hat{y} \equiv \hat{R}/\beta$, one obtains the following saddle point equations:

$$\begin{aligned} R &= 1 - 2H\left(\frac{\hat{y}}{\sqrt{\hat{c}}}\right) & c &= \sqrt{\frac{2}{\pi\hat{c}}}\exp\left(-\frac{\hat{y}^2}{2\hat{c}}\right) \\ \hat{c} &= \frac{\alpha}{c^2}\int \mathcal{D}t\, X(t;R)\left[\lambda_0(t,c) - t\right]^2 & \hat{y} &= \frac{\alpha}{c}\int \mathcal{D}t\, Y(t;R)\left[\lambda_0(t,c) - t\right]\ , \end{aligned} \tag{18}$$

where $\lambda_0(t,c) \equiv \mathrm{Argmin}_\lambda \left[V(\lambda) + (\lambda - t)^2/2c\right]$. The variational optimization of R with respect to the choice of V can now be performed as in refs. [11, 8, 12, 13] invoking the Schwarz inequality. We only quote the final result for the resulting overlap R_{opt} at the minimum of this optimal potential:

$$R_{opt} = 1 - 2H\left(\mathcal{F}(R_{opt})\right)\ . \tag{19}$$

The important issue to be examined is whether or not $R_{opt}(\alpha)$ saturates the bound given by the best binary. By comparison of eq. 19 with eq. 15, one immediately concludes that this is *not possible*, as long as $\mathcal{F}$ is not a constant nor singular, since $R_{opt} = R_{bb}$ would imply that $\mathcal{F}(R_{bb}) = \mathcal{F}(R_B)$, and $R_{bb} = R_B$ is excluded by the first equality in eq.15. In general one thus has that $R_{opt} \leq R_{bb}$. The equality is reached in asymptotic limits. For $R_{opt} \sim 0$ one has:

$$\langle t\rangle_* \neq 0 \quad \Rightarrow \quad R_{opt} \simeq |\langle t\rangle_*|\sqrt{\frac{2\alpha}{\pi}} \tag{20}$$

$$\langle t\rangle_* = 0 \quad \Rightarrow \quad R\begin{cases} = 0, & \alpha \leq \alpha_c \\ \simeq \sqrt{C'(\alpha - \alpha_c)}, & \alpha \geq \alpha_c \end{cases}\ , \tag{21}$$

where the critical value now is $\alpha_c \equiv \pi\alpha_G/2$. Furthermore, the approach $R_{opt} \to 1$ is identical to that of R_{bb}, $1 - R_{opt} \simeq 1 - R_{bb}$. Therefore V_{opt} is successful only in the asymptotic limits $\alpha \to 0$ and $\alpha \to \infty$. Note that the second order phase transition in eq. 21 occurs at a larger value of α than for Gibbs learning.

The case $\mathcal{F}(R)$ independent of R, implying $R_{opt} = R_{bb}$, $\forall\alpha$, arises in a simple Gaussian scenario with a linear function U [20]. In this case the best binary corresponds to clipped Hebbian learning. This seems to be the only case in which minimization of an optimal potential reproduces the best binary vector. This observation corroborates our feeling that potentials $V(\lambda)$ do not convey enough information to find an optimal vector in a binary space. It motivates the search for alternative methods in discrete optimization. The main issue is to find new ways to incorporate information about the binary nature of the symmetry breaking vector, other then simply imposing the same binary constraint in the solution space. An interesting approach would be to try to construct a suitable potential for the continuous center of mass $\mathbf{J}_B$ from which the best binary could be obtained from clipping. Whether such an approach is possible will be answered in future work.

***

The authors would like to thank Nestor Caticha for useful discussions and the organizers of the International Seminar on "Statistical Physics of Neural Networks" (1999), held at the Max Planck Institute for the Physics of Complex Systems (Dresden), during which part of this work was accomplished. We also acknowledge support from the FWO Vlaanderen and the Belgian IUAP program (Prime Minister's Office).

REFERENCES

[1] T. L. H. Watkin, A. Rau, and M. Biehl, *Rev. Mod. Phys.*, **65** (1993)

[2] M. Opper and W. Kinzel, in *Models of Neural Networks III*, edited by E. Domany, J. L. van Hemmen, and K. Schulten, (Springer-Verlag) 1996

[3] M. Biehl and A. Mietzner, *Europhys. Lett.*, **24** 421-426 (1993)

[4] M. Biehl and A. Mietzner, *J. Phys. A: Math. Gen.*, **27** 1885 (1994)

[5] T. L. H. Watkin and J.-P. Nadal, *J. Phys. A: Math. Gen.*, **27** 1899-1915 (1994)

[6] P. Reimann and C. Van den Broeck, *Phys. Rev. E*, **53** 3989 (1996)

[7] P. Reimann, C. Van de Broeck, and G. J. Bex, *J. Phys. A: Math. Gen.*, **29** 3521 (1996)

[8] C. Van den Broeck and P. Reimann, *Phys. Rev. Lett.*, **76** 2188 (1996)

[9] M. Bouten, J. Schietse and C. Van den Broeck, *Phys. Rev. E*, **52** 1958-1967 (1995)

[10] T. L. H. Watkin, *Europhys. Lett.*, **21** 871 (1993)

[11] O. Kinouchi and N. Caticha, *Phys. Rev. E*, **54** R54 (1996)

[12] A. Buhot, J.-M. Torres Moreno, and M. B. Gordon, *Phys. Rev. E*, **55** 7434-7440 (1997)

[13] A. Buhot and M. B. Gordon, *Phys. Rev. E*, **57** 3326-3333 (1998)

[14] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond*, World Scientific (Singapore) 1987

[15] G. Györgyi, *Phys. Rev. A*, **41** 7097 (1990)

[16] M. Opper and D. Haussler, *Phys. Rev. Lett.*, **66** 2677-2680 (1991)

[17] J. Schietse, M. Bouten, and C. Van den Broeck, *Europhys. Lett.*, **32** 279-284 (1995)

[18] Eq. 14 is consistent with the fact that the overlap for the center mass cannot be improved by a transformation of its components. Indeed the transformed overlap $\tilde{R}$ in eq. 11 is maximized [17] by setting $g_{opt}(x) = (P(x) - P(-x))/(P(x) + P(-x))$, reducing for (14) to $g_{opt}(x) \sim x$.

[19] C. R. de Mattos, *Aplicações de Mecânica Estatística ao Perceptron Binário e ao Processamento de Imagens,*, PhD thesis, Universidade de São Paulo (1997) (in Portuguese). Also N. Caticha, private communication.

[20] M. Copelli and C. Van den Broeck, unpublished