Retarded Learning: Rigorous Results from Statistical Mechanics

Didier Herschkowitz† Manfred Opper‡

† Laboratoire de Physique Statistique de L'E.N.S., Ecole Normale Supérieure, Paris,

‡ Neural Computing Research Group, Aston University, United Kingdom

August 5, 2000

Abstract

We study unsupervised learning of distributions which are characterized by a single unknown rotational-symmetry direction. We take the limit where the spatial dimension N and the number of data m grow large, keeping the ratio $\alpha = m/N$ fixed. In a Bayesian framework we develop upper and lower bounds on an entropic performance measure. We show that there is a critical ratio α_{lb} such that for $\alpha < \alpha_{lb}$ it is *impossible* at all to learn the distribution in the sense that the optimal Bayesian estimator performs as bad as a trivial estimator which always predicts an uniform distribution. On the

other hand we also give an upper bound on the critical ratio above which nontrivial learning is possible. The bounds allow to discuss the asymptotic behaviour of the learning for continuous and discontinuous distribution.

PACS numbers: 87.10.+e, 05.20.+m, 02.50.-r

In recent years, methods of Statistical Physics have contributed important insights into the theory of learning with neural networks and other learning machines (see e.g. [1, 2, 3]). Among the most prominent discoveries of statistical mechanics is the occurence of phase transitions in performance of learning, when data and parameters spaces of the model are high dimensional. This phenomenon describes a nonsmooth progress in the learning of an unknown rule, when the number of presented example data is gradually increased.

Besides the ubiquity of phase transitions in discrete parameter models, they are typically observed when the learning problem contains symmetries which are spontaneously broken when the scaled number of examples increases beyond a critical value [1, 4, 5]. Although phase transitions in neural networks have been analysed extensively by the method of replicas [6] it is usually hard to present a rigorous analysis (for an exception see e.g. [7] and the recent attempts of Talagrand [8]).

Hence, this often precludes a digestion of the interesting results by researchers outside the community of statistical physicists working on disordered systems.

Unfortunately, also other standard techniques based on asymptotic expansions [9] will not apply in these cases. They are only valid when the number of data is much larger than

the number of parameters.

In this letter we will present a rigorous but still simple approach to these problems. We combine information theoretic bounds for the performance of statistical estimators (see e.g. [10, 11, 12]) with an elementary variational principle of statistical physics [13]. This will allow us in many cases to compute rigorous upper and lower bounds for the critical number of examples at which a transition occurs.

We will explain our method for the case of retarded unsupervised learning which has been analysed before using the replica framework (see e.g. [14, 15, 16]). The goal of unsupervised learning is to find a nontrivial structure in a set of data which reflects the properties of the underlying data generating mechanism rather than being an artefact of statistical fluctuations. The phenomenon of retarded learning describes the fact that for some highly symmetric probability distributions, it is impossible at all to estimate the underlying structure (usually a symmetry direction) if the (scaled) number of data is below a certain critical value. Only above this value, estimation of the structure can start. A first attempt to study retarded learning by using exact information theoretic bounds was undertaken by [12]. However the bounds were too weak to give a nonzero bound on the critical number of examples below which learning is impossible. After introducing the cumulative risk in the general framework of unsupervised learning, we derive exact lower and upper bounds on this quantity and on the critical value after which learning can start. Then, we discuss the asymptotic behaviour of the learning for continuous and discontinuous distribution.

We adopt a probabilistic Bayesian formulation of unsupervised learning following [17, 10, 12]. We model a situation where the probability distribution of the data is characterized by a single unknown rotational-symmetry direction $\mathbf{w}^*$. More specifically, we assume that a set of t data $\boldsymbol{x}^t = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$, has been generated independently by t samplings from a distribution of the form

$$P(\boldsymbol{x}|\mathbf{w}^*) = P_0(\boldsymbol{x}) \ \exp\left(-V(\lambda)\right) \tag{1}$$

where $P_0(\boldsymbol{x}) = \frac{1}{(2\pi)^{N/2}} \exp(-\boldsymbol{x}^2/2)$ is a spherical Gaussian distribution and $\lambda \doteq \mathbf{w}^* \cdot \boldsymbol{x}$ is the projection of the data vector $\boldsymbol{x}$ on the direction defined by $\mathbf{w}^*$. The distribution of the projection is given by $p(\lambda) = \frac{1}{\sqrt{2\pi}} \exp(-\lambda^2/2 - V(\lambda))$. In the following, averaged quantities with respect to this distribution will be denoted with an overline $\overline{(..)} = \int d\lambda p(\lambda)(..)$. Based on the set of data $\boldsymbol{x}^t$, the goal of a learner is to produce an estimation $\hat{P}_t(\boldsymbol{x}|\boldsymbol{x}^t)$ for the true distribution $P(\boldsymbol{x}|\mathbf{w}^*)$. $\hat{P}_t$ will not necessarily belong to the given parametric class (1). The expected quality of the estimation can be measured by the averaged *relative entropy* (Kullback Leibler (KL) divergence) between the true distribution and the estimation

$$L(\hat{P}_t, \mathbf{w}^*) \doteq \int d\boldsymbol{x}^t P(\boldsymbol{x}^t|\mathbf{w}^*) \int d\boldsymbol{x} P(\boldsymbol{x}|\mathbf{w}^*) \ln \frac{P(\boldsymbol{x}|\mathbf{w}^*)}{\hat{P}_t(\boldsymbol{x}|\boldsymbol{x}^t)} \tag{2}$$

where $P(\boldsymbol{x}^t|\mathbf{w}^*)$ is shortand for the product distribution $\prod_{i=1}^t P(\boldsymbol{x}_i|\mathbf{w}^*)$ and $d\boldsymbol{x}^t = \prod_{i=1}^m d\boldsymbol{x}_i$. We will further adopt a Bayesian approach where we assume that "nature" draws the true parameter $\mathbf{w}^*$ at random from a noninformative prior distribution $p(\mathbf{w}^*)$ and associate measure $d\mu(\mathbf{w}^*) = p(\mathbf{w}^*)d\mathbf{w}^*$ given by the uniform distribution on the sphere with radius $\|\mathbf{w}^*\|^2 = 1$. The case of a discrete prior will be discussed at the end.

The progress of the learning will be measured by the cumulative risk defined by

$$R_m(\hat{P}) = \sum_{t=0}^{m-1} \int d\mu(\mathbf{w}^*) L(\hat{P}_t, \mathbf{w}^*) \tag{3}$$

This measure of loss in the game of estimation is often easy to evaluate and has a variety of important applications in information theory, game theory and mathematical finance (see e.g. [10]) and also in statistical physics, as we will see in a moment. E.g. it yields a measure for the expected redundancy in compressing the training data when the true distribution is unknown and a sequential estimate is used instead for encoding. An elementary calculation shows that the posterior probability $\hat{P}_t^{Bayes}(\boldsymbol{x}|\boldsymbol{x}^t) = \frac{\int d\mu(\mathbf{w})\ P(\boldsymbol{x}^t|\mathbf{w})P(\boldsymbol{x}|\mathbf{w})}{\int d\mu(\mathbf{w}')\ P(\boldsymbol{x}^t|\mathbf{w}')}$ achieves the *minimum* risk $R_m^{Bayes} = R_m(\hat{P}^{Bayes})$ over all choices of estimators. Inserting this estimator into (3) and using (1) we get

$$R_m^{Bayes} \;=\; \int d\mu(\mathbf{w}^*) \int d\boldsymbol{x}^m P(\boldsymbol{x}^m|\mathbf{w}^*) \left[-\ln \int d\mu(\mathbf{w}) e^{-\sum_i \{V(\mathbf{w}\cdot\boldsymbol{x}_i) - V(\mathbf{w}^*\cdot\boldsymbol{x}_i)\}} \right] \tag{4}$$

The last line looks very much like an averaged free energy in statistical mechanics for a system with hamiltonian $\sum_i \{V(\mathbf{w}\cdot\boldsymbol{x}_i) - V(\mathbf{w}^*\cdot\boldsymbol{x}_i)\}$. Hence we can expect that useful bounds for this quantity can be derived using the standard variational principle of statistical mechanics [13] for the free energy.

$$-\ln \int d\mu(\mathbf{w}) e^{-H(\mathbf{w})} \leq -\ln \int d\mu(\mathbf{w}) e^{-H_0(\mathbf{w})} + \langle H - H_0 \rangle_0 \tag{5}$$

which bounds the free energy of a system with hamiltonian H in terms of the free energy of a trial hamiltonian H_0 plus a correction term. The brackets $\langle (..) \rangle_0$ =

$\int d\mu(\mathbf{w}) e^{-H_0(\mathbf{w})}(..) / \int d\mu(\mathbf{w}') e^{-H_0(\mathbf{w}')}$ denote an average with respect to the Gibbs distribution defined by H_0.

We set $H_0 = \sum_i \{V(\mathbf{w} \cdot \boldsymbol{x}_i) - V(\mathbf{w}^* \cdot \boldsymbol{x}_i)\}$ and $H = \sum_i \{\lambda V(\mathbf{w} \cdot \boldsymbol{x}_i) - \gamma V(\mathbf{w}^* \cdot \boldsymbol{x}_i)\}$ where $\lambda, \gamma > 0$ are variational parameters. Averaging over $P(\boldsymbol{x}^m|\mathbf{w}^*)$ and $p(\mathbf{w}^*)$ both sides of (5) using Jensen's inequality in the second line, we derive the lower bound on R_m^{Bayes}

$$\begin{aligned} R_m^{Bayes} &\geq \int d\mu(\mathbf{w}^*) d\boldsymbol{x}^m P(\boldsymbol{x}^m|\mathbf{w}^*) \left[-\ln \int d\mu(\mathbf{w}) e^{-H} \right] + m(\gamma - \lambda)\overline{V} \\ &\geq -\int d\mu(\mathbf{w}^*) \ln \int d\mu(\mathbf{w}) \left[\int d\boldsymbol{x} P_0(\boldsymbol{x}) \frac{e^{-\lambda V(\boldsymbol{x}\cdot\mathbf{w})}}{e^{(1-\gamma)V(\boldsymbol{x}\cdot\mathbf{w}^*)}} \right]^m + m(\gamma - \lambda)\overline{V} \\ &= -\ln \left\{ \int_{-1}^{1} dq \; W_N(q) \left[F_{\lambda\gamma}(q) \right]^m \right\} + m \; (\gamma - \lambda)\overline{V} \end{aligned}$$

where $W_N(q) = \int d\mu(\mathbf{w}) \; \delta(q - \mathbf{w}.\mathbf{w}^*) \propto (1 - q^2)^{\frac{N-3}{2}}$ and

$$F_{\lambda\gamma}(q) = \int Dx \int Dy e^{-\lambda V(x) - (1-\gamma) V(xq + y\sqrt{1-q^2})}$$

$Dx = e^{-\frac{1}{2}x^2} dx / \sqrt{2\pi}$ is the gaussian measure. This bound holds for every N and every m.

For convenience, we will compare the cumulative risk of the Bayes estimator to the risk of a *trivial estimator* which assumes that there is no specific structure in the data and always predicts with the spherical distribution $\hat{P}_t^{triv}(\boldsymbol{x}) = P_0(\boldsymbol{x})$ thereby achieving the trivial total risk $R_m^{triv} = R_m(\hat{P}^{triv}) = -m\overline{V}$. Note, that $\overline{V}$ is a nonpositive quantity. We are interested in the difference $\Delta R_m = R_m^{triv} - R_m^{Bayes}$ between the trivial and Bayes risk. In order to show the phenomenon of retarded learning, we take the thermodynamic limit $N \to \infty$ with $\alpha = m/N$ fixed. The integral can be evaluated by Laplace's method which gives an upper bound for

ΔR_m

$$\lim_{N\to\infty} \Delta R_{\alpha N}/N \;\; \leq \;\; \min_q \left\{ \frac{1}{2} \ln(1-q^2) + \alpha \ln F_{\gamma\lambda}(q) \right\} + \alpha(\lambda - \gamma - 1)\overline{V} \tag{6}$$

For sufficiently small α, the bound (6) is optimized for $\gamma = 0$ and $\lambda = 1$. For any potential V having the property $\overline{\lambda} = 0$ (ie. when the problem is not trivially learnable by computing the mean of the data) there is some critical value $\alpha_{lb} = (1 - \overline{\lambda^2})^{-2}$ such that as long as $\alpha \leq \alpha_{lb}$ the minmizer is $q = 0$ and $\lim_{N\to\infty} \Delta R_{\alpha N}/N \;\; \leq 0$ (see Fig. 1). As the Bayes risk is optimal $\Delta R_m \geq 0$ and we conclude that $\lim_{N\to\infty} \Delta R_{\alpha N}/N \;\; = 0$ at least for $\alpha \leq \alpha_{lb}$. This proves the existence of a region of retarded learning, where even the risk of the optimal Bayes estimator is to leading order in N as large as the risk of a trivial estimator which assumes that there is no any spatial structure at all. Hence, it is also *impossible* to estimate the rotational-symmetry direction of the data generating distribution.

We now derive a lower bound on ΔR_m. Using the fact that R_m^{Bayes} is the minimimum cumulative risk over any choice of estimators, for any distribution $Q(\boldsymbol{x}|\mathbf{w})$ and estimator $\hat{Q}(\boldsymbol{x}|\boldsymbol{x}^t) = \frac{\int d\mu(\mathbf{w}) \;\; Q(\boldsymbol{x}^t|\mathbf{w}) Q(\boldsymbol{x}|\mathbf{w})}{\int d\mu(\mathbf{w}') \;\; Q(\boldsymbol{x}^t|\mathbf{w}')}$ we have $R_m^{Bayes} \leq R_m(\hat{Q})$. Setting $H = -\ln \frac{Q(\boldsymbol{x}^m|\mathbf{w})}{P(\boldsymbol{x}^m|\mathbf{w}^*)}$ in (5) and restricting ourselves to the class of trial Hamiltonians H_0 which do not depend on $\boldsymbol{x}^m$, it can be shown that the optimal choice is the data average $H_0 = \int d\boldsymbol{x}^m \;\; P(\boldsymbol{x}^m|\mathbf{w}^*) \;\; H$, for which, on average, the correction term in (5) vanishes. This yields

$$\begin{aligned} R_m^{Bayes} \;\; &\leq \;\; R_m(\hat{Q}) \\ &= \;\; -\int d\mu(\mathbf{w}^*) d\boldsymbol{x}^m P(\boldsymbol{x}^m|\mathbf{w}^*) \ln \frac{\int d\mu(\mathbf{w}) \;\; Q(\boldsymbol{x}^m|\mathbf{w})}{P(\boldsymbol{x}^m|\mathbf{w}^*)} \end{aligned} \tag{7}$$

$$\leq \; -\int d\mu(\mathbf{w}^*) \ln \int d\mu(\mathbf{w}) \exp -m \int d\boldsymbol{x} P(\boldsymbol{x}|\mathbf{w}^*) \ln \frac{Q(\boldsymbol{x}|\mathbf{w})}{P(\boldsymbol{x}|\mathbf{w}^*)}$$

We now have to find a good choice for $Q(\boldsymbol{x}|\mathbf{w})$. For $Q(\boldsymbol{x}|\mathbf{w})$ has a structure of the form (1), we set $Q(\boldsymbol{x}|\mathbf{w}) = P_0(\boldsymbol{x}) \; \exp\left(-U(\mathbf{w} \cdot \boldsymbol{x})\right)$ such that in the thermodynamic limit we get the lower bound for ΔR_m

$$\lim_{N\to\infty} \Delta R_{\alpha N}/N \geq \max_q \left\{ \frac{1}{2} \ln(1-q^2) \; -\alpha \int Dx \exp(-U_q(x)) U(x) \right\} \tag{8}$$

with

$$U_q(x) = -\ln \int Dy \exp -V(xq + y\sqrt{1-q^2}). \tag{9}$$

It is easy to see that the bound is maximised for $U(x) = U_q(x)$. With this choice for U, we find that there exists an α_{ub} such that for $\alpha > \alpha_{ub}$, we have $\lim_{N\to\infty} \Delta R_{\alpha N}/N \; > 0$ which means that now the performance of the Bayes risk is better than the trivial risk and a nontrivial estimation of the direction $\mathbf{w}^*$ is possible (see Fig. 1). α_{ub} is an upper bound on the region of retarded learning but has no simple analytical expression.

We now discuss the asymptotic behaviour $\alpha \to \infty$ for R_m^{Bayes}/N expanding the bounds (6) and (8) for $q \to 1$. For a smooth potential V both bounds give asymptotically the same logarithmic growth $R_m^{Bayes}/N \to 1/2 \; ln \; \alpha$ which can also be obtained by well known asymptotc expansions involving the *Fisher information matrix* [18, 19, 11]. However, when the potential exhibits a discontinuity of the form $V(\lambda) = -\ln \; 2\Theta(\lambda) + U(\lambda)$ with corresponding projection distribution $p(\lambda) = \Theta(\lambda) \frac{2}{\sqrt{2\pi}} \exp(-\lambda^2/2 - U(\lambda))$ where $\Theta(\lambda)$ is the Heaviside function and U is smooth, the standard asymptotic expansions do not apply. However, our

bounds yield $R_m^{Bayes}/N \to \ ln\ \alpha$.

In virtue of (7) our bounds are also bounds on the perfomance of the estimator $\hat{Q}$ which uses the optimizing potential U_q. It follows that $R_m(\hat{Q})$ has the same asymptotic behavior as the optimal R_m^{Bayes}. This could simplify the estimation of discontinuous probability in using smoothened distribution like (9) without loosing much in perfomance asymptotically. For example, following (9), the case $V(\lambda) = -\ln 2\Theta(\lambda)$ can be well estimated using $U_q(\lambda) = -\ln\ 2H(-q\lambda/\sqrt{1-q^2})$ where $H(x) = \int_x^\infty Dx$ and where q solution of (8), is an adjustable parameter for each value of α.

Our approach is also easily applied to a discrete prior distribution, e.g. a uniform distribution on the hypercube [20]. Again, for small α we find a region of retarded learning $\Delta R_m^{Bayes}/N = 0$ at least for $\alpha \leq \alpha_{lb}$ where α_{lb} is exactly the same as for the spherical prior.

For illustration, we apply our bounds to the simple case of a Gaussian distribution for which all integrals can be done analytically. Other distributions will be discussed in [21]. We set $P(\boldsymbol{x}|\mathbf{w}^*) = \frac{1}{(2\pi)^{N/2}(1+A)} \exp\left(-\frac{\boldsymbol{x}^2}{2} + \frac{A}{2(1+A)}(\boldsymbol{x}\cdot\mathbf{w}^*)^2\right)$. The data are normally distributed with unit variance in all directions perpendicular to $\mathbf{w}^*$ and with variance $1+A$ in the direction $\mathbf{w}^*$. The upper and lower bounds (6) and (8) for $N \to \infty$ (optmized w.r.t. λ and γ) are shown in Figure 1 for $A = -0.5$ for which we obtain $\alpha_{lb} = 1/A^2 = 4$. We have compared our bounds with numerical simulations. Since it is hard to compute the Bayes optimal estimation algorithmically, we have used the following (suboptimal) algorithm instead. From the data set $\boldsymbol{x}^t$ generated by $P(\boldsymbol{x}|\mathbf{w}^*)$, we have computed the estimated direction $\hat{\mathbf{w}}(\boldsymbol{x}^t)$ for

$\mathbf{w}^*$ in a way similar to [14], using a maximum likelihood algorithm $\hat{\mathbf{w}} = arg\max_{\mathbf{w}} \ln P(\boldsymbol{x}^t|\mathbf{w})$. We then estimate the true distribution using the plugin estimator $\hat{P}_t^p(\boldsymbol{x}|\boldsymbol{x}^t) \doteq P(\boldsymbol{x}|\hat{\mathbf{w}}(\boldsymbol{x}^t))$ which gives the KL divergence

$$\int d\boldsymbol{x} P(\boldsymbol{x}|\mathbf{w}^*) \ln \frac{P(\boldsymbol{x}|\mathbf{w}^*)}{\hat{P}_t^p(\boldsymbol{x}|\boldsymbol{x}^t)} = \frac{A^2}{1+A}(1 - (\mathbf{w}^* \cdot \hat{\mathbf{w}}(\boldsymbol{x}^t))^2)$$

and the cumulative entropic risk $R_m(\hat{P}^p)$ can be easily approximated numerically by averaging over a large number of data sets. Figure 1 shows the difference $R_m(P^{triv}) - R_m(\hat{P}^p)$. As we can see, until $\alpha \approx \alpha_{lb}$, we are in the retarded learning regime: the error of the plugin estimator is even worse than the trivial estimator. This is due to the fact that the plugin

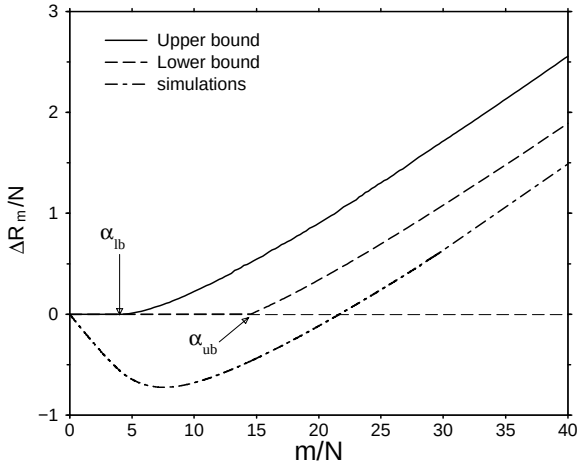

Figure 1: Upper and lower bound for the Bayes Risk $\Delta R_m/N = (R_m^{Bayes} - R_m^{trivial})/N$ in the limit $N \to \infty$ for the gaussian case $A = -0.5$. $\alpha_{lb} = 4$. Numerical simulation of $(R_m(P^{triv}) - R_m(\hat{P}^p))/N$ with $N = 100$ for the plugin estimator average over 50 data sets.

estimator has to keep an ellipsoidal form which is always different from spherical in a region

where no special direction appears. As soon as nontrivial learning of $\mathbf{w}^*$ becomes possible, the performance of the plugin estimator change to become better than the trivial one. Then the curve start to increase. The Bayes estimator does not have this kind of disadvantage and can take a form closer to the spherical distribution in the retarded learning region by smoothing over the parameters $\mathbf{w}^*$. Remark that as the Bayes risk is optimal, the upper bound on ΔR_m is also an upper bound on every estimator while the lower bound is only a lower bound on the Bayes risk.

It is interesting to note that the plugin estimator $\hat{P}_t^p(\boldsymbol{x}|\boldsymbol{x}^t) \doteq P_0(\boldsymbol{x}) \exp(-U_q(\boldsymbol{x}|\hat{\mathbf{w}}(\boldsymbol{x}^t)))$ where U_q is given by (9) and with $q = \mathbf{w}^* \cdot \hat{\mathbf{w}}$ would have minimized the loss (2) and improved the learning performances over any other choices of potential in the exponent (see [21]). $q = \mathbf{w}^* \cdot \hat{\mathbf{w}}$ is not known in general, but it would be calculable in the thermodynamic limit with replica technique.

In this contribution, we have put the phenomenon of retarded learning first established by the replica method of statistical mechanics on a rigorous footing, by optimization of information theoretic risk bounds. These bounds were derived by a very simple variational method. The quality of such bounds extend from a region where the number of data is much larger than the dimension of the parameter space to a nonasymptotic region where it is difficult to obtain rigorous results.

Our bounds and the critical value α_{lb} as a lower bound for the region of retarded learning are in very good agreement with results obtained with the statistical mechanics "replica"

technique [15, 20, 12] and in very good agreement with numerical simulations. We also proved that estimation of discontinuous distribution using smoothened effective distribution of the type (9) can lead to asymptotic performance close to optimum.

Our bounds apply to very differents situation like spherical and discrete priors and discontinuous likelihoods for which standard asymptotic expansions can not be applied. Although we have worked with a simple model class with only a single rotational-symmetry direction, we expect that the effect of retarded learning should also be relevant in realistic situations. It would be important to further develop a theory which can be helpful in practice to decide if structures that have been estimated from a dataset in a high dimensional space reflects a real feature of the underlying distribution or if the result is expected to be a spurious effect of random fluctuations.

We are grateful to J.-P. Nadal for helpful discussions.

References

[1] H.S. Seung, H. Sompolinsky and N. Tishby, *Phys. Rev. A*, 45: 6056-6091, 1992.

[2] T. L. H. Watkin, A. Rau and M. Biehl; Rev. Mod. Phys. 65, 499 (1993).

[3] M. Opper and W. Kinzel; *Statistical Mechanics of Generalization*, *Models of Neural Networks III*, ed. by J. L. van Hemmen, E. Domany and K. Schulten, published by Springer Verlag (1996).

[4] H. Shwarze, M. Opper and W. Kinzel, *Phys. Rev. A*, 46:R6185-R6188, 1992.

[5] M. Opper, *Phys. Rev. Lett.*, 72:2113-2116, 1994; *Phys. Rev. E*, 51 (4):3613-3618, 1995.

[6] M. Mezard, G. Parisi and M. A. Virasoro, *Spin glass theory and beyond*, World Scientific, Singapore, 1987

[7] D. Haussler, M. Kearns, S. Seung and N. Tishby, Machine Learning 25(2-3):195-236, 1996.

[8] http://www.math.ohio-state.edu/ talagran/

[9] S. Amari and N. Murata, *Neural Computation*, 5:140-153, 1993.

[10] D. Haussler and M. Opper, *Annals of Statistics* 25 (6):2451-2492, 1997;

[11] N. Brunel and J.-P. Nadal, *Neural Computation*, 10 (7):1731-1757, 1998.

[12] D. Herschkowitz and J.-P. Nadal, *Phys. Rev. E*, 59 (3):3344-3360, 1999.

[13] R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals*, McGraw-Hill, Inc., 1965

[14] M. Biehl and A. Mietzner, *Europhys. Lett.*, 24 (5):421-426, 1993.

[15] P. Reimann and C. Van den Broeck, *Phys. Rev. E*, 53 (4):3989-3998, 1996.

[16] A. Buhot and M. Gordon, *Phys. Rev. E*, 57(3):3326-3333, 1998.

[17] M. Opper and D. Haussler, *Phys. Rev. Lett.* 75:3772-3775, 1995.

[18] B. S. Clarke and A. R. Barron,

IEEE Trans. on Information Theory, 36 (3):453-471, 1990.

[19] J. Rissanen, *IEEE Trans. on Information Theory*, 42 (1):40-47, 1996.

[20] M. Copelli, C. Van den Broeck and M. Opper *J. Phys. A*, 32: L555-L560, 1999.

[21] D. Herschkowitz, Ph.D. thesis, 2000

[22] M. B. Gordon and D. Grempel *Europhys. Lett.*, 29 (3): 257-262, 1995.