Retarded Learning: Rigorous Results from Statistical Mechanics

Didier Herschkowitz[1] and Manfred Opper[2]

[1]*Laboratoire de Physique Statistique de L'E.N.S., Ecole Normale Supérieure, Paris, France*
[2]*Neural Computing Research Group, Aston University, United Kingdom*

We study learning of probability distributions characterized by an unknown symmetry direction. Based on an entropic performance measure and the variational method of statistical mechanics we develop exact upper and lower bounds on the scaled critical number of examples below which learning of the direction is impossible. The asymptotic tightness of the bounds suggests an asymptotically optimal method for learning nonsmooth distributions.

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., [1–3]). Among the most prominent discoveries of statistical mechanics in this field is the occurrence of phase transitions in the progress of learning, when the number of presented example data is gradually increased and the dimensionality of data and model parameter is large.

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 analyzed 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 valid only when the number of data is much larger than the number of parameters.

In this Letter we will present a rigorous and simple approach to these problems. We combine information theoretic bounds for the performance of statistical estimators (see, e.g., [10–12]) with an elementary variational principle of statistical physics [13]. This will allow us 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 analyzed before using the replica framework (see, e.g., [14–17]). 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 artifact of statistical fluctuations. The phenomenon of retarded learning describes the fact that for some high dimensional probability distributions, it is impossible at all to predict the underlying structure (usually a symmetry axis) if the

(scaled) number of data is below a certain critical value. Only above this value, estimation of the structure can start.

We adopt a probabilistic, Bayesian formulation of unsupervised learning following [10,15,18]. 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

$$P(\boldsymbol{x}|\mathbf{w}^*) = P_0(\boldsymbol{x}) \exp[-V(\lambda)], \qquad (1)$$

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 N-dimensional 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 $p(\lambda)$ will be denoted with an overline $\overline{(\cdots)} = \int d\lambda\, p(\lambda)\, (\cdots)$. Based on the set of data $\boldsymbol{x}^t$, the goal of a learner is to produce an estimate $\hat{P}_t(\boldsymbol{x}|\boldsymbol{x}^t)$ for the true distribution $P(\boldsymbol{x}|\mathbf{w}^*)$. $\hat{P}_t$ will not necessarily belong to the given parametric class (1). Our approach relies very much on the choice of a specific measure for the quality of the estimation. Rather than computing the overlap between an estimated direction and the true $\mathbf{w}^*$, we choose a quantity which is directly measuring our ability to compress the data based on the information we have gained on the structure of P. This is related to the averaged *relative entropy* [Kullback Leibler (KL) divergence] between the true distribution and the estimate

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

where $P(\boldsymbol{x}^t|\mathbf{w}^*)$ is shorthand for the product distribution $\prod_{i=1}^{t} P(\boldsymbol{x}_i|\mathbf{w}^*)$ and $d\boldsymbol{x}^t = \prod_{i=1}^{t} d\boldsymbol{x}_i$. 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 later.

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

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

This measure of loss has a variety of important applications in information theory, game theory, and mathematical finance (see, e.g., [10,19]). For example, it is proportional to the expected extra number of bits (assuming a reasonable quantization of the $\mathbf{x}_i$) we have to suffer in compressing the observed data when their distribution is unknown and a sequential estimate is used instead [19]. As we will see in a moment it has also an important meaning in statistical physics.

A first attempt to study retarded learning by using bounds on (3) 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.

An elementary calculation shows that the posterior probability $\hat{P}_t^{\mathrm{Bayes}}(\mathbf{x}|\mathbf{x}^t) = \frac{\int d\mu(\mathbf{w})P(\mathbf{x}^t|\mathbf{w})P(\mathbf{x}|\mathbf{w})}{\int d\mu(\mathbf{w}')P(\mathbf{x}^t|\mathbf{w}')}$ achieves the *minimum* risk $R_m^{\mathrm{Bayes}} = R_m(\hat{P}^{\mathrm{Bayes}})$ over all choices of estimators. Inserting this estimator into (3) and using (1) we get

$$R_m^{\mathrm{Bayes}} = \int d\mu(\mathbf{w}^*) \int d\mathbf{x}^m P(\mathbf{x}^m|\mathbf{w}^*)$$

$$\times \left[-\ln \int d\mu(\mathbf{w})e^{-\sum_i\{V(\mathbf{w}\cdot\mathbf{x}_i)-V(\mathbf{w}^*\cdot\mathbf{x}_i)\}} \right]. \qquad (4)$$

The last line looks very much like an averaged free energy in statistical mechanics for a system with Hamiltonian $\sum_i\{V(\mathbf{w}\cdot\mathbf{x}_i) - V(\mathbf{w}^*\cdot\mathbf{x}_i)\}$. 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

$$-\ln \int d\mu(\mathbf{w})e^{-H(\mathbf{w})} \le -\ln \int d\mu(\mathbf{w})e^{-H_0(\mathbf{w})}$$

$$+ \langle H - H_0 \rangle_0 \qquad (5)$$

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 $\langle(\cdots)\rangle_0 = \int d\mu(\mathbf{w})e^{-H_0(\mathbf{w})}(\cdots)/\int d\mu(\mathbf{w}')e^{-H_0(\mathbf{w}')}$ denote an average with respect to the Gibbs distribution defined by H_0. Using appropriate choices for H and H_0, we will get both upper and lower bounds on (4).

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

$$R_m^{\mathrm{Bayes}} \ge \int d\mu(\mathbf{w}^*) d\mathbf{x}^m P(\mathbf{x}^m|\mathbf{w}^*)\left[-\ln \int d\mu(\mathbf{w})e^{-H} \right] + m(\gamma - \lambda)\overline{V}$$

$$\ge -\int d\mu(\mathbf{w}^*)\ln \int d\mu(\mathbf{w})\left[\int d\mathbf{x}P_0(\mathbf{x})\frac{e^{-\lambda V(\mathbf{x}\cdot\mathbf{w})}}{e^{(1-\gamma)V(\mathbf{x}\cdot\mathbf{w}^*)}} \right]^m + m(\gamma - \lambda)\overline{V}$$

$$= -\ln\left\{ \int_{-1}^{1} dq W_N(q)[F_{\lambda\gamma}(q)]^m \right\} + m(\gamma - \lambda)\overline{V}, \qquad (6)$$

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

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

with the Gaussian measure $Dx = e^{-(1/2)x^2}dx/\sqrt{2\pi}$. This bound holds for every N and every m.

To show the phenomenon of retarded learning 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 $\hat{P}_t^{\mathrm{triv}}(\mathbf{x}) = P_0(\mathbf{x})$ thereby achieving the trivial total risk $R_m^{\mathrm{triv}} = R_m(\hat{P}^{\mathrm{triv}}) = -m\overline{V}$. Note, that $\overline{V}$ is a nonpositive quantity. We are interested in the difference $\Delta R_m = R_m^{\mathrm{triv}} - R_m^{\mathrm{Bayes}}$ between the trivial risk and the Bayes risk. Taking the thermodynamic limit $N \to \infty$ with $\alpha = m/N$ fixed, we can evaluate the integral in (6) by Laplace's method which gives the asymptotic upper bound for ΔR_m

$$\limsup_{N\to\infty}\Delta R_{\alpha N}/N \le \min_q\left\{ \frac{1}{2}\ln(1 - q^2) + \alpha \ln F_{\gamma\lambda}(q) \right\}$$

$$+ \alpha(\lambda - \gamma - 1)\overline{V}. \qquad (7)$$

For sufficiently small α, the bound (7) is optimized for $\gamma = 0$ and $\lambda = 1$. For any potential V having the property $\overline{\lambda} = 0$ (i.e., when the problem is not trivially learnable by computing the mean of the data) there is a critical value $\alpha_{lb} = (1 - \overline{\lambda^2})^{-2}$ such that as long as $\alpha \le \alpha_{lb}$ the minimizer is $q = 0$ and $\lim_{N\to\infty}\Delta R_{\alpha N}/N \le 0$ (see Fig. 1). Since the Bayes risk is minimal, we have $\Delta R_m \ge 0$ and we conclude that $\lim_{N\to\infty}\Delta R_{\alpha N}/N = 0$ at least for $\alpha \le \alpha_{lb}$. 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 spatial structure at all. The bound α_{lb} agrees with the critical α obtained in the replica analysis of [15].

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

$$R_m^{\text{Bayes}} \leq R_m(\hat{Q})$$
$$= -\int d\mu(\mathbf{w}^*) \, dx^m P(x^m|\mathbf{w}^*) \ln\frac{\int d\mu(\mathbf{w})Q(x^m|\mathbf{w})}{P(x^m|\mathbf{w}^*)}$$
$$\leq -\int d\mu(\mathbf{w}^*) \ln \int d\mu(\mathbf{w}) \exp\left[-m \int dx P(x|\mathbf{w}^*) \ln\frac{Q(x|\mathbf{w})}{P(x|\mathbf{w}^*)}\right]. \qquad (8)$$

We now have to find a good choice for $Q(x|\mathbf{w})$. For $Q(x|\mathbf{w})$ with a structure of the form (1), we set $Q(x|\mathbf{w}) = P_0(x)\exp[-U(\mathbf{w}\cdot x)]$. In the thermodynamic limit we get the lower bound for ΔR_m

$$\liminf_{N\to\infty}\Delta R_{\alpha N}/N \geq \max_q\left\{\frac{1}{2}\ln(1-q^2) - \alpha\int Dx\,\exp[-U_q(x)]U(x)\right\} \qquad (9)$$

with

$$U_q(x) = -\ln\int Dy\,\exp[-V(xq + y\sqrt{1-q^2})]. \quad (10)$$

It is easy to see that for any q, the expression in the curly brackets of (9) is maximized 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} gives an upper bound on the region of retarded learning but has no simple analytical expression.

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
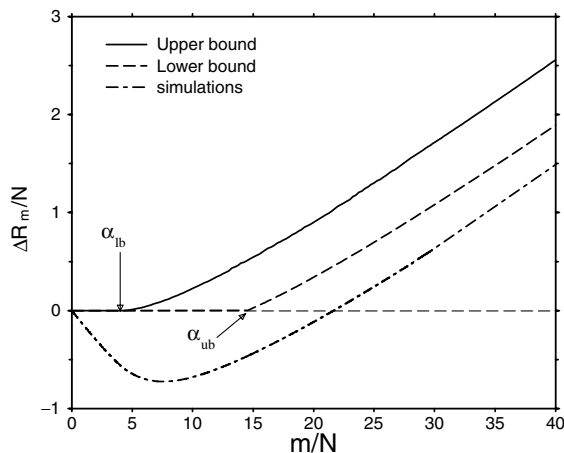

FIG. 1. Upper and lower bound for the Bayes Risk $\Delta R_m/N = (R_m^{\text{Bayes}} - R_m^{\text{trivial}})/N$ in the limit $N \to \infty$ for the Gaussian case $A = -0.5$. $\alpha_{lb} = 4$. Simulations are for the plug-in estimator and show $[R_m(P^{\text{triv}}) - R_m(\hat{P}^p)]/N$ for $N = 100$ averaged over 50 data sets.

learning $\Delta R_m^{\text{Bayes}}/N = 0$ at least for $\alpha \leq \alpha_{lb}$ where α_{lb} is exactly the same as for the spherical prior.

For illustration, we apply the bounds to the simple case of a Gaussian distribution for which the integrals can be done analytically. Other distributions will be discussed in [21]. We set $P(x|\mathbf{w}^*) = \frac{1}{(2\pi)^{N/2}(1+A)}\exp[-\frac{x^2}{2} + \frac{A}{2(1+A)}(x\cdot\mathbf{w}^*)^2]$. 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 (7) and (9) (optimized with respect to λ and γ) are shown in Fig. 1 for $A = -0.5$ for which we obtain $\alpha_{lb} = 1/A^2 = 4$. We have compared the bounds with numerical simulations. Since it is hard to compute the Bayes optimal estimation algorithmically, we have used the following (suboptimal) algorithm instead. We have computed the direction $\hat{\mathbf{w}}(x^t)$ for $\mathbf{w}^*$ which maximizes the posterior probability of the data. The estimate of the distribution is given by the plug-in estimator $\hat{P}_t^p(x|x^t) \doteq P(x|\hat{\mathbf{w}}(x^t))$ which has the KL divergence

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

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 $R_m(P^{\text{triv}}) - R_m(\hat{P}^p)$. Since the Bayes risk is minimal, 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. We see that until $\alpha \approx 4$, ΔR_m is negative and decreases linearly. Since the slope of the dash-dotted curve is proportional to the (relative) *instantaneous* loss (i.e., the noncumulative risk), the plug-in estimator is in the retarded learning regime; its instantaneous loss is even bigger than that of the trivial estimator.

This is due to the fact that the plug-in estimator has to keep the elliptic form of the distribution $P(\boldsymbol{x}|\mathbf{w})$ which is always different from spherical when $\hat{\mathbf{w}}(\boldsymbol{x}^t) \neq 0$. Between $\alpha \geq 4$ and $\alpha \leq 7.5$, the performance of the plug-in estimator improves, the slope increases but is still negative. For $\alpha > 7.5$, the performance of the plug-in estimator is better than the trivial one and the curve starts 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 $\mathbf{w}^*$.

Both bounds in Fig. 1 show the same type of asymptotic growth for $\alpha \to \infty$. The asymptotics can be found analytically by expanding both bounds (7) and (9) for $q \to 1$. For a smooth potential V both bounds give asymptotically the same logarithmic growth $R_m^{\text{Bayes}}/N \simeq 1/2\ln\alpha$ which can also be obtained by well known asymptotic expansions involving the *Fisher information matrix* [11,22,23]. On the other hand, our bounds can also be used when these standard asymptotic expansions do not apply, e.g., when the potential exhibits a discontinuity [15] of the form $V(\lambda) = -\ln2\Theta(\lambda) + U(\lambda)$ 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, Our bounds yield the asymptotic scaling $R_m^{\text{Bayes}}/N \simeq \ln\alpha$.

The asymptotic matching of our bounds has an important consequence for computing asymptotically good approximations to Bayesian predictions. Such approximations are easily derived for smooth potentials by local expansions of the posterior distribution around its maximum (Ref. [24]). However, this technique will obviously fail for nonsmooth potentials. On the other hand, our results show that the estimate $\hat{Q}$ in (8) which uses the *smooth* optimizing potential U_q (10), has the same asymptotic performance as the Bayes optimal estimate R_m^{Bayes}. The smoothness of U_q will again enable local expansions. For example, following (10), the case $V(\lambda) = -\ln2\Theta(\lambda)$ can be well estimated using $U_q(\lambda) = -\ln2H(-q\lambda/\sqrt{1 - q^2})$ with $H(x) = \int_x^\infty Dx$ and where the maximizer q of the right hand side of (9) is a function of α.

In this Letter, we have put the phenomenon of retarded learning first established by the replica method on a rigorous footing. Our method relies on a general information theoretic performance measure for learning probability distributions which is related to the free energy of statistical physics. A variational principle yields a controlled approximation to this quantity by providing exact upper and lower bounds which are valid for arbitrary dimensionality of the problem. We expect that this framework is flexible enough to be applicable to more complex and realistic probabilistic models. It may also be useful for constructing criteria that help to decide if structures estimated from a dataset in a high dimensional space reflect a real feature of the underlying data generating mechanism or if the result is expected to be a spurious effect of random fluctuations.

[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,* edited by J. L. van Hemmen, E. Domany, and K. Schulten (Springer-Verlag, Berlin, 1996).

[4] H. Schwarze, 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**, 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, Mach. Learn. **25**, 195–236 (1996).

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

[9] S. Amari and N. Murata, Neural Comput. **5**, 140–153 (1992).

[10] D. Haussler and M. Opper, Ann. Stat. **25**, 2451–2492 (1997).

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

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

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

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

[15] P. Reimann and C. Van den Broeck, Phys. Rev. E **53**, 3989–3998 (1996); C. Van den Broeck and P. Reimann, Phys. Rev. Lett. **76**, 2188–2191 (1996).

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

[17] B. Schottky, J. Phys. A **28**, 4515–4531 (1995).

[18] M. Opper and D. Haussler, Phys. Rev. Lett. **75**, 3772–3775 (1995).

[19] T. Cover and J. A. Thomas, *Element of Information Theory,* Wiley Series in Telecommunications (Wiley, New York, 1991).

[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] B. S. Clarke and A. R. Barron, IEEE Trans. Inf. Theory **36**, 453–471 (1990).

[23] J. Rissanen, IEEE Trans. Inf. Theory **42**, 40–47 (1996).

[24] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis* (Springer-Verlag, New York, 1985).

[25] O. Kinouchi and N. Caticha, J. Phys. A **25**, 6243 (1992).

[26] M. B. Gordon and D. Grempel, Europhys. Lett. **29**, 257–262 (1995).