Chapter 20

Maximum Entropy Connections

Maximum entropy estimation of probability distributions constitutes a theoretical foundation for the Hopfield associative memory. Subject to knowledge of the first and second order statistics of a collection of binary variables, the maximum entropy distribution is the exponential of a Hopfield network's energy. In a special case, an explicit expression for the connection strengths in this maximum entropy net is given; this converges exactly to Hopfield's covariance prescription in the limit of large numbers of random patterns. This connection between probabilistic inference and neural networks gives a viewpoint on the effective assumptions and approximations being made when a Hopfield network is used as an associative memory, and motivates several modifications to the original algorithms.

This early work discusses a probabilistic view of a neural network learning algorithm that involves density modelling. It does not fall into the Bayesian framework described in the rest of this book, but it may be of interest all the same.

20.1 The Hopfield associative memory

Architecture

I will describe the simplest Hopfield network (43). The network consists of N 'neurons' which have binary activities $x_i = \pm 1$. The neurons are pairwise connected through symmetrical weights $w_{ij} = w_{ji}$. The neurons are not connected to themselves, so $w_{ii} = 0$.

Dynamics

The dynamics of an individual neuron x_i depend on the value of its activation a_i, which is a linear combination of the other x_j:

$$a_i = \sum_j w_{ij} x_j + \theta_i \tag{20.1}$$

where θ_i is the bias of neuron i. The simplest update rule is to change the x_i asynchronously according to:

$$x_i(t+1) = \begin{cases} 1 & \text{if } a_i(t) > 0 \\ -1 & \text{if } a_i(t) < 0 \end{cases} . \tag{20.2}$$

Energy function

The Hopfield network has a Lyapunov function

$$E = -\frac{1}{2}\sum_{i,j} x_i w_{ij} x_j - \sum_i \theta_i x_i, \tag{20.3}$$

which is minimized by the dynamics (18.2). The network may be used as an associative memory, if there is a prescription for the parameters such that the minima of the energy function can be located at an arbitrary set of memory states $\mathbf{x}$.

Prescription for connection strengths

Given a set of random binary patterns $\{\mathbf{x}^{(m)}\}$, $m = 1, 2 ... \alpha N$, which are to be stored as memories in the network, the covariance prescription for connection strengths is ( )

$$w_{ij} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle,$$

where the averages are taken over the set of memory vectors. The hope is that, given this choice of the parameters, the dynamics of the network will converge from any starting point to the nearest vector in the list of memories $\{\mathbf{x}^{(m)}\}$. The capacity of the network has been studied using spin glass techniques (4), and it has been shown that this prescription stores random patterns successfully (with less than 1.5% errors) for α up to ~ 0.14. Higher capacities can be achieved by more complex algorithms (31).

20.2 Maximum Entropy inference of probability distributions

Consider a world of variables $\mathbf{x}$ on which there is an unknown probability distribution $p(\mathbf{x})$. Let the average value of various functions $f_i(\mathbf{x})$ be observed, implying a set of constraints on $p(\mathbf{x})$:

$$\int f_i(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \bar{f}_i \tag{20.4}$$

This information only gives us partial knowledge about $p(\mathbf{x})$. There are generally many probability distributions satisfying the constraints. The Maximum Entropy principle provides a criterion for choosing from those valid probability distributions a unique preferred $p(\mathbf{x})$ which is maximally non–committal, including only the information provided by the constraints. Maxent chooses the $p(\mathbf{x})$ that maximizes the entropy

$$S(p(\mathbf{x})) = -\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}$$

subject to the constraints (18.4). Introducing a Lagrange multiplier α_i for each constraint and differentiating, the maximum entropy probability distribution has the form

$$p(\mathbf{x}) = \exp\left(\sum_i \alpha_i f_i(\mathbf{x})\right), \tag{20.5}$$

where the Lagrange multipliers are fixed by the constraints (18.4). The constraint that $p(\mathbf{x})$ should be normalized introduces a factor of e^{α_0} in (18.5).

20.3 Maximum entropy associative memory

In a binary world $\mathbf{x}$, assume that we have knowledge of the average value of each variable $\langle x_i \rangle$, and of the average pairwise cross–correlation $\langle x_i x_j \rangle$ between each pair of variables, or equivalently the covariance $c_{ij} \equiv \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$. This knowledge constitutes 'testable information' of the form (18.4). Let us assume we have no other prior knowledge (note this means that we explicitly neglect the knowledge that the true probability distribution over the world may actually be a set of αN delta functions located at the 'memories'). Then the maximum entropy probability distribution, from (18.5), has the form:

$$p(\mathbf{x}) = \exp\left(\alpha_0 + \sum_i \theta_i x_i + \frac{1}{2}\sum_{i,j} w_{ij} x_i x_j\right) \tag{20.6}$$

where θ_i and w_{ij} are Lagrange multipliers yet to be determined. Comparing this expression with (18.3), we see that the maximum entropy probability distribution is exactly the exponential of (minus) a Hopfield network's energy, so that that Hopfield network's dynamics find local maxima of the maximum entropy probability distribution.

The question now to be addressed is how the parameters of the network are set by maximum entropy. There is not a closed form solution to this general problem. The solution to any instance of a maximum entropy problem can be found iteratively using the 'Boltzmann Machine' neural network ( ): if $p^*(\mathbf{x})$ is the true probability distribution over $\mathbf{x}$ and $p(\mathbf{x})$ is a Boltzmann distribution $p(\mathbf{x}) = \exp\left(\sum_i \alpha_i f_i(\mathbf{x})\right)$, the Boltzmann Machine's objective function $G = p^*(\mathbf{x}) \log p(\mathbf{x})/p^*(\mathbf{x})$ has its maximum exactly when the Lagrange multipliers α_i solve the maximum entropy problem defined in section 18.2. In the special case that follows however, an explicit maximum entropy solution is possible.

Maximum entropy connections

Consider a set of binary variables x_i which are the conditioning 'inputs' to a single binary 'output' variable y. Assume the average values $\langle x_i \rangle$ and $\langle y \rangle$ have been measured, and the pairwise cross–correlations $\langle x_i y \rangle$ between each x_i and y (but not among the x_i). Then the maximum entropy distribution over $(\mathbf{x}, y)$ is:

$$p(\mathbf{x}, y) = \exp\left(\alpha_0 + \theta_y y + \sum_i w_{yi} x_i y + \theta_i x_i\right) \tag{20.7}$$

In the language of neural networks, the output variable y corresponds to a single neuron which receives connections from a collection of input neurons x_i. Note that under the maximum entropy distribution (18.7), the variables x_i are independent given y, *i.e.* we can write: $P(y|\{x_i\}) \propto \prod_i P(x_i|y)P(y)$. This means that our neuron is a linear Bayes classifier, and the values of w_{yi} and θ_y can be derived by examining Bayes' rule and identifying the parameters with the appropriate log conditional probabilities: The log of the posterior probability ratio is:

$$\ln \frac{P(y=1|\{x_i\})}{P(y=-1|\{x_i\})} = \sum_i \ln \frac{P(x_i|y=1)}{P(x_i|y=-1)} + \ln \frac{P(y=1)}{P(y=-1)} \tag{20.8}$$

We compare this with the log probability ratio from the maximum entropy expression (18.7).

$$\ln \frac{P(y=1|\{x_i\})}{P(y=-1|\{x_i\})} = \sum_i 2 w_{yi} x_i + 2\theta_y \tag{20.9}$$

Note in passing that this log probability ratio is equal to twice a_i, the activation (equation 18.1). The aim is now to put (18.8) into the form $\sum_i 2w_{yi}x_i+2\theta_i$, and evaluate the maximum entropy parameters w_{yi} in terms of the statistics of the memory vectors. To do this, we rewrite the term inside the summation as a linear function of x_i and obtain:

$$w_{yi} = \frac{1}{4}\ln\frac{P(x_i=1,y=1)P(x_i=-1,y=-1)}{P(x_i=1,y=-1)P(x_i=-1,y=1)} \tag{20.10}$$

and

$$\theta_y = \frac{1}{2}\ln\frac{P(y=1)}{P(y=-1)} + \frac{1}{4}\sum_i \ln\frac{P(x_i=1|y=1)P(x_i=-1|y=1)}{P(x_i=1|y=-1)P(x_i=-1|y=-1)}. \tag{20.11}$$

The arguments thus far can be found in (, 40). It may be confirmed that the remaining parameters in (18.7) are:

$$\theta_i = \frac{1}{4}\ln\frac{P(x_i=1,y=1)P(x_i=1,y=-1)}{P(x_i=-1,y=-1)P(x_i=-1,y=1)}.$$

Comparison with covariance

We now compare the expression for w_{yi} (18.10) with Hopfield's covariance expression by substituting for the covariance $c_{yi} = \langle yx_i\rangle - \langle y\rangle\langle x_i\rangle$.

Letting $p_i = P(x_i=1)$, $q_i = P(x_i=-1)$, $p_y = P(y=1)$ and $q_y = P(y=-1)$, and using relationships of the form $P(y=1,x_i=1) = p_yp_i + \frac{1}{4}c_{yi}$, we obtain:

$$w_{yi} = \frac{1}{4}\ln\frac{\left(1+\frac{c_{yi}/4}{p_yp_i}\right)\left(1+\frac{c_{yi}/4}{q_yq_i}\right)}{\left(1-\frac{c_{yi}/4}{q_yp_i}\right)\left(1-\frac{c_{yi}/4}{p_yq_i}\right)} \tag{20.12}$$

Equation (18.12) is still exact. Now if we assume that the bits are almost independent, for example if the statistics are derived from a large number of random patterns, so that the terms $\frac{c_{yi}/4}{p_yp_i}$, etc. are small compared to 1, then we can Taylor expand and obtain in the weak covariance limit:

$$w_{yi} \simeq \frac{1}{4}\left(\frac{1}{p_yp_i}+\frac{1}{q_yq_i}+\frac{1}{q_yp_i}+\frac{1}{p_yq_i}\right)\frac{c_{yi}}{4} = \frac{c_{yi}}{16\,p_yp_iq_yq_i} \tag{20.13}$$

Thus if $p_i = \beta\,\forall i$, the maximum entropy connection strengths are to first order proportional to the covariance between units. In the case $p_i = \frac{1}{2}\,\forall i$, the exact expression (18.12) can be massaged into a more familiar form:

$$w_{yi} = \frac{1}{4}\ln\frac{(1+c_{yi})(1+c_{yi})}{(1-c_{yi})(1-c_{yi})} = \frac{1}{2}\ln\frac{(1+c_{yi})}{(1-c_{yi})} = \tanh^{-1} c_{yi} \tag{20.14}$$

The first order approximation to this expression is $w_{yi} = c_{yi}$, exactly as prescribed by Hopfield. But it is interesting to note that the exact expression assigns stronger than linear weights to strong covariances. This is intuitive, since if there is a perfect correlation between two variables ($c_{yi} \to 1$) then there should be an infinite weight between them to maintain that correlation.

Thus maximum entropy provides a derivation of the covariance–based associative memory. This corroborates the work of Willshaw and Dayan (112), who use a signal to noise criterion to derive the covariance learning rule as the optimal local rule for hetero–associative memory.

20.4 New ideas provoked by this connection

Alternative expressions for w_{ij}

The expressions for w_{ij} derived thus far apply only to a special case in which the correlations between the set of input variables were not measured. In the general case of a fully connected network where strong inter–variable correlations are measured, the maximum entropy solution for the parameters will certainly differ from the covariance prescription. However, if the statistics are generated from random patterns, it is conjectured that $w_{ij} = c_{ij}$ remains the first order expression for the maximum entropy connections. The previous section motivates modifications to this covariance prescription, which might prove useful in cases of intermediate strength correlations; simple alternative expressions have been given in equations 18.10, 18.13 and 18.14.

Expression for bias

In the case where the bit probabilities p_i are not 0.5, the maximum entropy connection motivates an expression for the bias (18.11), the leading term of which is:

$$\theta_i = \ln \sqrt{\frac{p_i}{q_i}}. \tag{20.15}$$

The use of non–zero biases is discussed in (89); how (18.15) relates to that work has not been established.

Comments on the properties of the covariance prescription

It was conjectured above that when the statistics are generated from random patterns, $w_{ij} = c_{ij}$ is the correct first order expression for the maximum entropy connections. If this conjecture is correct then some interesting insights follow.

First, the so–called 'spurious states' at which local minima of E appear, although those states were not in the memory set, should not be viewed as meaningless artifacts of an imperfect memory storage algorithm. Such states have been *inferred* to be probable states by maximum entropy's generalization from the measured statistics.

Second, the breakdown of the Hopfield network as an associative memory above $\alpha \simeq 0.14$ is not due to a poor approximation of maximum entropy. Indeed, as the number of random patterns stored increases, the covariances decrease as $1/\sqrt{\alpha N}$, so the covariance prescription is expected to approach the maximum entropy solution more closely as α increases. Rather, this breakdown is due to the fact that the Hopfield net is actually solving a different problem from memorization of a list of patterns. Recall from section 18.3 that the maximum entropy derivation rested on ignoring the prior knowledge that the true probability distribution is a list of delta functions. Thus the memory works below $\alpha \simeq 0.14$ only because of the happy coincidence that the maxima of the maximum entropy distribution happen to coincide with the memories that generated the statistics. Under this interpretation, what the work of Amit et. al. shows is that above $\alpha \simeq 0.14$, the maxima of the maximum entropy distribution no longer coincide with those patterns — but the Hopfield network still infers the correct maximum entropy probability distribution, given the statistics it was provided with.

Choice of temperature and gain

Instead of the deterministic dynamics (18.2), stochastic dynamics are frequently considered. These generate network states $\mathbf{x}$ according to a Boltzmann distribution $p(\mathbf{x}) =$

$\exp -E(\mathbf{x})/T$. The maximum entropy equation for $p(\mathbf{x})$ (18.6) contains no temperature variable. Assuming again that the covariance prescription is the maximum entropy solution for w_{ij}, this motivates the special status of a temperature of 1. It is interesting to note that this is the temperature above which the stochastic network's dynamics are paramagnetic, *i.e.* the temperature above which the energy minima are no longer attracting ( ).

Similarly, when discrete activities are replaced by continuous valued activities representing probabilities or mean field values (83), a unique value for the gain for the sigmoid transfer function is motivated. This gain will be $O(1)$ (the precise value will depend on the details of the implementation).

Alternative procedure for memory recall from the Hopfield dynamics

Thus far this paper has related the Hopfield network to maximum entropy solely through its energy function, and little reference has been made to the dynamics that should be used to access the knowledge stored in a maximum entropy memory.

I will now suggest a procedure for memory recall that is different from the Hopfield dynamics. Assume that we are provided with a cue vector that is a corrupted version of a memory from the original statistics. Hopfield's suggested recall procedure was to set the network off in the cue state and then let the dynamics (18.2) take over. This is not a Bayesian inference procedure, and it leads to problems which indicate that improvements could be made. For example, if the memory is overloaded (*i.e.* more than $0.14N$ patterns are stored), then the initial cue vector is 'forgotten' and the dynamics take the state of the network to an energy minimum unrelated to the cue. This is clearly not a sensible inference.

Bayesian inference about the current state of the world $\mathbf{x}$ is based on evaluating the posterior probability distribution, which is shaped by two competing forces: the cue vector, which gives partial information about the current state $\mathbf{x}$; and the prior knowledge about the statistics of the patterns, which is embodied in the net's energy function. The log posterior is the sum of the log prior and the log likelihood term, which is separable into a sum over the variables if the noise on each cue bit is independent.

$$\log P(\mathbf{x}|\mathrm{cue}) = -E(\mathbf{x}) + \sum_i \log P(\mathrm{cue}_i|x_i)$$

The term inside the summation can be written as a linear function of x_i. Thus the noisy cue manifests itself as a linear bias added to the energy $E(\mathbf{x})$. If the bit transition probability between x_i and cue_i is b_i ($0 < b_i < \frac{1}{2}$; the bigger b_i is, the less reliable cue_i is), then the strength of this 'applied field' is found to be:

$$\theta'_i = \mathrm{cue}_i \frac{1}{2} \log \frac{1-b_i}{b_i}. \tag{20.16}$$

If associative memory is viewed as finding the $\mathbf{x}$ with maximum posterior probability, this motivates the use of the network dynamics (18.2) with the noisy cue applied as a *sustained bias throughout recall.*

As shown in equation (18.16), this Bayesian approach allows the reliability of the cue to be represented quantitatively, since the strength of this bias is related to how noisy each bit of the cue is.

The idea of presenting a sustained bias during memory recall has recently appeared in the literature, without this Bayesian motivation (27 , 5 , 115). Such procedures have been shown to enhance memory recall.

Generalization to networks with higher than pairwise connectivity.

The expression for w_{ij} (18.10) generalizes to networks with connections between more than two neurons (such networks are discussed in (91)). For example, the analogous prescription for a third–order connection is

$$\gamma_{ijk} = \frac{1}{8} \ln \frac{P(+++)P(+--)P(-+-)P(--+)}{P(++-)P(+-+)P(-++)P(---)}$$

where $P(-++) = P(x_i = -1, x_j = 1, x_k = 1)$, etc.

20.5 Comments

The maximum entropy derivation of the Hopfield energy function assumed that we were given hard constraints of the form $\langle x_i \rangle = \mu_i$. However many applications will supply us with limited amounts of data such that we have inexact knowledge of $\langle x_i \rangle$ and $\langle x_i x_j \rangle$. Hard constraints only result in the limit of an infinite amount of data or in cases where a symmetry property provides a prior constraint. For clarity I have omitted this complication from this paper. This issue can be confronted head on with a full Bayesian analysis, in which the entropy becomes the log prior, and we examine the posterior distribution over $p(\mathbf{x})$:

$$P(p(\mathbf{x})|\text{Data}) \propto P(\text{Data}|p(\mathbf{x}))e^{\alpha S(p(\mathbf{x}))}$$

However for practical purposes it is probably adequate just to take the expressions derived assuming a hard constraint and substitute in orthodox 'robust estimators' of the covariances.

Bibliography

[1] Y.S. Abu-Mostafa. Learning from hints in neural networks. *J. Complexity*, 6:192–198, 1990.

[2] Y.S. Abu-Mostafa. The Vapnik–Chervonenkis dimension: information versus complexity in learning. *Neural Computation*, 1(3):312–317, 1990.

[3] H. Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217, 1970.

[4] D.J. Amit, H. Gutfrend, and H. Sompolinsky. Statistical mechanics of neural networks near saturation. *Ann. Phys. (New York)*, 173:30, 1987.

[5] D.J. Amit, G. Parisi, and S. Nicolis. Neural potentials as stimuli for attractor neural networks. *Network*, 1(1):75–88, 1990.

[6] L.R. Bahl, F. Jelinek, and R.L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Trans*, PAMI–5(2):179–190, 1983.

[7] P. Baldi and Heiligenberg. How sensory maps could enhance resolution through ordered arrangement of broadly tuned receivers. *Biol. Cyb.*, 59:313–318, 1988.

[8] Y. Bar-Shalom and T.E. Fortmann. *Tracking and Data Association.* Academic Press, 1988.

[9] E.B. Baum. Neural net algorithms that learn in polynomial time from examples and queries. *IEEE Trans. on neural networks*, 2(1):5–19, 1991.

[10] J. Berger. *Statistical Decision theory and Bayesian Analysis.* Springer, 1985.

[11] C.M. Bishop. Exact calculation of the Hessian matrix for the multilayer perceptron. *Neural Computation*, 4(4):494–501, 1992.

[12] G.E.P. Box and G.C. Tiao. A further look at robustness via Bayes' theorem. *Biometrika*, 49:419–432, 1962.

[13] G.E.P. Box and G.C. Tiao. A Bayesian approach to some outlier problems. *Biometrika*, 55:119–129, 1968.

[14] G.E.P. Box and G.C. Tiao. *Bayesian inference in statistical analysis.* Addison–Wesley, 1973.

[15] L. Breiman. Stacked regressions. Technical Report 367, Dept. of Stat., Univ. of Cal. Berkeley, 1992.

[16] G.L. Bretthorst. *Bayesian spectrum analysis and parameter estimation.* Springer, 1988.

[17] J.S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. Fougelman-Soulie and editors J. Hérault, editors, *Neuro-computing: algorithms, architectures and applications.* Springer–Verlag, 1989.

[18] R.K. Bryan. Solving oversampled data problems by Maximum Entropy. In P. Fougere, editor, *Maximum Entropy and Bayesian Methods, Dartmouth, U.S.A., 1989*, pages 221–232. Kluwer, 1990.

[19] W.L. Buntine and A.S. Weigend. Bayesian back–propagation. *Complex Systems*, 5:603–643, 1991.

[20] D.H. Wolpert C.E.M. Strauss and D.R. Wolf. Alpha, evidence, and the entropic prior. In A. Mohammed-Djafari, editor, *Maximum Entropy and Bayesian Methods, Paris 1992*, Dordrecht, 1993. Kluwer.

[21] M.K. Charter. Quantifying drug absorption. In W.T. Grandy and L. Schick, editors, *Maximum Entropy and Bayesian Methods, Laramie, 1990*, pages 245–252, Dordrecht, 1991. Kluwer.

[22] J.B. Copas. Regression, prediction and shrinkage (with discussion). *J.R.Statist.Soc B*, 45(3):311–354, 1983.

[23] R.T. Cox. Probability, frequency, and reasonable expectation. *Am. J. Physics*, 14:1–13, 1946.

[24] A.R. Davies and R.S Anderssen. Optimization in the regularization of ill–posed problems. *J. Austral. Mat. Soc. Ser. B*, 28:114–133, 1986.

[25] J.S. Denker and Y. LeCun. Transforming neural-net output levels to probability distributions. In R.P. Lippmann, editor, *Advances in Neural Information Processing Systems 3*, pages 853–859. Morgan Kaufmann, San Mateo, California, 1991.

[26] M.A. El-Gamal. The role of priors in active Bayesian learning in the sequential statistical decision framework. In W.T. Grandy and L. Schick, editors, *Maximum Entropy and Bayesian Methods, Laramie, 1990*, pages 33–38, Dordrecht, 1991. Kluwer.

[27] A. Engel, H. English, and A. Schutte. Improved retrieval in neural networks with external fields. *Europhys Lett.*, 8:393, 1989.

[28] Rumelhart *et al.*. *Parallel Distributed Processing.* MIT Press, 1986.

[29] R.L. Eubank. *Spline smoothing and non–parametric regression.* Marcel Dekker, 1988.

[30] V.V. Fedorov. *Theory of optimal experiments.* Academic press, 1972.

[31] E.J. Gardner. Maximum storage capacity of neural networks. *Europhys. Lett.*, 4:481, 1987.

[32] S.F. Gull. Bayesian inductive inference and maximum entropy. In G.J. Erickson and C.R. Smith, editors, *Maximum Entropy and Bayesian Methods in Science and Engineering, vol. 1: Foundations*, pages 53–74, Dordrecht, 1988. Kluwer.

[33] S.F. Gull. Bayesian data analysis: straight–line fitting. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods, Cambridge 1988*, pages 511–518, Dordrecht, 1989. Kluwer.

[34] S.F. Gull. Developments in maximum entropy data analysis. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods, Cambridge 1988*, pages 53–71, Dordrecht, 1989. Kluwer.

[35] S.F. Gull and G.J. Daniell. Image reconstruction from incomplete and noisy data. *Nature*, 272:686–690, 1978.

[36] I. Guyon, V.N. Vapnik, B.E. Boser, L.Y. Bottou, and S.A. Solla. Structural risk minimization for character recognition. In J.E. Moody, S.J. Hanson, and R.P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 471–479, San Mateo, California, 1992. Morgan Kaufmann.

[37] R. Hanson, J. Stutz, and Peter Cheeseman. Bayesian classification theory. Technical Report FIA–90-12-7-01, NASA Ames, 1991.

[38] B. Hassibi and D.G. Stork. Second order derivatives for network pruning: Optimal brain surgeon. In C.L. Giles, S.J. Hanson, and J.D. Cowan, editors, *Advances in Neural Information Processing Systems 5*, pages 164–171, San Mateo, California, 1993. Morgan Kaufmann.

[39] D. Haussler, M. Kearns, and R. Schapire. Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. In *Proceedings of the fourth COLT workshop*, San Mateo, California, 1991. Morgan Kaufmann.

[40] G.E. Hinton and T.J. Sejnowski. Optimal perceptual inference. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 448–453, 1983.

[41] G.E. Hinton and T.J. Sejnowski. Learning and relearning in boltzmann machines. In Rumelhart *et al.*, editor, *Parallel Distributed Processing*, pages pp. 282–317. MIT Press, 1986.

[42] G.E. Hinton and D. van Camp. Keeping neural networks simple by minimizing the description length of the weights. To appear in: *Proceedings of COLT-93*, 1993.

[43] J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, 79:2554–8, 1982.

[44] J.J. Hopfield. Learning algorithms and probability distributions in feed–forward and feed–back networks. *Proc. Natl. Acad. Sci. USA*, 84:8429–33, 1987.

[45] J-N. Hwang, J.J. Choi, S. Oh, and R.J. Marks II. Query–based learning applied to partially trained multilayer perceptrons. *IEEE Trans. on Neural networks*, 2(1):131–136, 1991.

[46] E.T. Jaynes. Bayesian intervals versus confidence intervals. In R.D. Rosencrantz, editor, *E.T. Jaynes. Papers on Probability, Statistics and Statistical Physics*, page 151. Kluwer, 1983.

[47] E.T. Jaynes. Bayesian methods: general background. In J.H. Justice, editor, *Maximum Entropy and Bayesian Methods in applied statistics*, pages 1–25. C.U.P., 1986.

[48] H. Jeffreys. *Theory of Probability.* Oxford Univ. Press, 1939.

[49] C. Ji, R.R. Snapp, and D. Psaltis. Generalizing smoothness constraints from discrete samples. *Neural Computation*, 2(2):188–197, 1990.

[50] Y. LeCun, J.S. Denker, and S.A. Solla. Optimal brain damage. In D.S. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, pages 598–605. Morgan Kaufmann, 1990.

[51] W.T. Lee and M.F. Tenorio. On optimal adaptive classifier design criterion — how many hidden units are necessary for an optimal neural network classifier? Technical Report TR-EE-91-5, Purdue University, 1991.

[52] E. Levin, N. Tishby, and S.A. Solla. A statistical approach to learning and generalization in layered neural networks. In *COLT '89: 2nd workshop on computational learning theory*, pages 245–260, 1989.

[53] D.V. Lindley. On a measure of the information provided by an experiment. *Ann. Math. Statist.*, 27:986–1005, 1956.

[54] D.V. Lindley. Bayesian analysis in regression problems. In D.L. Meyer and eds. R.O. Collier, editors, *Bayesian statistics.* Peacock publishers, 1970.

[55] D.V. Lindley. *Bayesian statistics, a review.* Society for Industrial and Applied Mathematics, Philadelphia, 1972.

[56] T.J. Loredo. From Laplace to supernova SN 1987A: Bayesian inference in astrophysics. In P. Fougere, editor, *Maximum Entropy and Bayesian Methods, Dartmouth, U.S.A., 1989*, pages 81–142. Kluwer, 1990.

[57] S.P. Luttrell. The use of transinformation in the design of data sampling schemes for inverse problems. *Inverse Problems*, 1:199–218, 1985.

[58] S.P. Luttrell. The use of Bayesian and entropic methods in neural network theory. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods, Cambridge 1988*, pages 363–370, Dordrecht, 1989. Kluwer.

[59] D.J.C. MacKay. *Bayesian Methods for Adaptive Models.* PhD thesis, California Institute of Technology, 1991.

[60] D.J.C. MacKay. Maximum entropy connections: Neural networks. In W.T. Grandy and L. Schick, editors, *Maximum Entropy and Bayesian Methods, Laramie, 1990*, Dordrecht, 1991. Kluwer.

[61] D.J.C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[62] D.J.C. MacKay. Bayesian model comparison and backprop nets. In J.E. Moody, S.J. Hanson, and R.P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 839–846, San Mateo, California, 1992. Morgan Kaufmann.

[63] D.J.C. MacKay. The evidence framework applied to classification networks. *Neural Computation*, 4(5):698–714, 1992.

[64] D.J.C. MacKay. Information based objective functions for active data selection. *Neural Computation*, 4(4):589–603, 1992.

[65] D.J.C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

[66] D.J.C. MacKay. Bayesian methods for backpropagation networks. In J.L. van Hemmen, E. Domany, and K. Schulten, editors, *Models of Neural Networks II*. Springer-Verlag, New York, 1993.

[67] D.J.C. MacKay. Bayesian non-linear modeling for the 1993 energy prediction competition. Technical Report in preparation, University of Cambridge, 1993.

[68] D.J.C. MacKay. Hyperparameters: Optimise, or integrate out? In G. Heidbreder, editor, *Maximum Entropy and Bayesian Methods, Santa Barbara 1993*, Dordrecht, 1994. Kluwer.

[69] D.J.C. MacKay and R.M. Neal. Automatic relevance determination for neural networks. Technical Report in preparation, Cambridge University, 1993.

[70] J.D. Mollon and J.K. Bowmaker. The spatial arrangement of cones in the primate fovea. *Nature*, 360:677–679, 1992.

[71] J.E. Moody. Note on generalization, regularization and architecture selection in nonlinear learning systems. In *First IEEE–SP Workshop on neural networks for signal processing*, pages 847–854. IEEE Computer society press, 1991.

[72] J.E. Moody. The *effective* number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In J.E. Moody, S.J. Hanson, and R.P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 847–854, San Mateo, California, 1992. Morgan Kaufmann.

[73] A. Nadas. Estimation of probabilities in the language model of the IBM speech recognition system. *IEEE Trans*, ASSP–32(4):859–861, 1984.

[74] R.M. Neal. Bayesian mixture modeling by Monte Carlo simulation. Technical Report Preprint, Dept. of Computer Science, University of Toronto, 1991.

[75] R.M. Neal. Bayesian training of backpropagation networks by the hybrid Monte Carlo method. Technical Report CRG–TR–92–1, Dept. of Computer Science, University of Toronto, 1992.

[76] R.M. Neal. Bayesian learning via stochastic dynamics. In C.L. Giles, S.J. Hanson, and J.D. Cowan, editors, *Advances in Neural Information Processing Systems 5*, pages 475–482, San Mateo, California, 1993. Morgan Kaufmann.

[77] R.M. Neal. Priors for infinite networks. Technical Report in preparation, Univ. of Toronto, 1993.

[78] R.M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG–TR–93–1, Dept. of Computer Science, University of Toronto, 1993.

[79] S.J. Nowlan. *Soft competitive adaptation: neural Network learning algorithms based on fitting statistical mixtures*. PhD thesis, Carnegie Mellon University, 1991. CS–91–126.

[80] S.J. Nowlan and G.E. Hinton. Adaptive soft weight tying using Gaussian mixtures. In J.E. Moody, S.J. Hanson, and R.P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 993–1000, San Mateo, California, 1992. Morgan Kaufmann.

[81] D.B. Osteyee and I.J. Good. *Information, weight of evidence, the singularity between probability measures and signal detection.* Springer, 1974.

[82] J.D. Patrick and C.S. Wallace. Stone circle geometries: an information theory approach. In D.C. Heggie, editor, *Archaeoastronomy in the Old World*, pages 231–264. Cambridge Univ. Press, 1982.

[83] C. Peterson and J.R. Anderson. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019, 1987.

[84] F.J. Pineda. Recurrent back–propagation and the dynamical approach to adaptive neural computation. *Neural Computation*, 1:161–172, 1989.

[85] M. Plutowski and H. White. Active selection of training examples for network learning in noiseless environments. Technical Report TR 90-011, Dept. Computer Science, UCSD, 1991.

[86] T. Poggio, V. Torre, and C. Koch. Computational vision and regularization theory. *Nature*, 317(6035):314–319, 1985.

[87] W.H. Press, S.A. Teukolsky B.P. Flannery, and W.T. Vetterling. *Numerical Recipes in C.* Cambridge, 1988.

[88] F. Reif. *Fundamentals of statistical and thermal physics.* McGraw–Hill, 1965.

[89] N. Rubin and H. Sompolinsky. Neural networks with low local firing rates. *Europhys. Lett.*, 8:465, 1989.

[90] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning representations by back–propagating errors. *Nature*, 323:533–536, 1986.

[91] T.J. Sejnowski. Higher order boltzmann machines. In J.S. Denker, editor, *Neural networks for computing*, pages 398–403, New York, 1986. American Institute of Physics.

[92] H.S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. Technical report, preprint, 1991.

[93] S. Sibisi. Bayesian interpolation. In W.T. Grandy and L. Schick, editors, *Maximum Entropy and Bayesian Methods, Laramie, 1990*, pages 349–355, Dordrecht, 1991. Kluwer.

[94] J. Skilling. Classic maximum entropy. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods, Cambridge 1988*, Dordrecht, 1989. Kluwer.

[95] J. Skilling. The eigenvalues of mega–dimensional matrices. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods, Cambridge 1988*, pages 455–466, Dordrecht, 1989. Kluwer.

[96] J. Skilling. On parameter estimation and quantified maxent. In W.T. Grandy and L. Schick, editors, *Maximum Entropy and Bayesian Methods, Laramie, 1990*, pages 267–273, Dordrecht, 1991. Kluwer.

[97] J. Skilling. Bayesian solution of ordinary differential equations. In G.J. Erickson and C.R. Smith, editors, *Maximum Entropy and Bayesian Methods, Seattle 1991,*, Dordrecht, 1992. Kluwer.

[98] J. Skilling. Bayesian numerical analysis. In W.T. Grandy Jr. and P. Milonni, editors, *Physics and Probability*, Cambridge, 1993. C.U.P.

[99] J. Skilling, D.R.T. Robinson, and S.F. Gull. Probabilistic displays. In W.T. Grandy and L. Schick, editors, *Maximum Entropy and Bayesian Methods, Laramie, 1990*, pages 365–368, Dordrecht, 1991. Kluwer.

[100] A.F.M. Smith and D.J. Spiegelhalter. Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society B*, 42(2):213–220, 1980.

[101] S.A. Solla, E. Levin, and M. Fleisher. Accelerated learning in layered neural networks. *Complex systems*, 2:625–640, 1988.

[102] D.J. Spiegelhalter and S.L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:579–605, 1990.

[103] S.M. Stigler. Laplace's 1774 memoir on inverse probability. *Stat. Sci.*, 1(3):359–378, 1986.

[104] H.H. Thodberg. Ace of Bayes: application of neural networks with pruning. Technical Report 1132 E, Danish meat research institute, 1993.

[105] N. Tishby, E. Levin, and S.A. Solla. Consistent inference of probabilities in layered networks: predictions and generalization. In *Proc. IJCNN, Washington*, 1989.

[106] D. Titterington. Common structure of smoothing techniques in statistics. *Int. Statist. Rev.*, 53:141–170, 1985.

[107] A.M. Walker. On the asymptotic behaviour of posterior distributions. *J. R. Stat. Soc. B*, 31:80–88, 1967.

[108] C.S. Wallace and D.M. Boulton. An information measure for classification. *Comput. J.*, 11(2):185–194, 1968.

[109] C.S. Wallace and P.R. Freeman. Estimation and inference by compact coding. *J.R. Statist. Soc. B*, 49(3):240–265, 1987.

[110] A.S. Weigend, D.E. Rumelhart, and B.A. Huberman. Generalization by weight–elimination with applications to forecasting. In R.P. Lippmann et. al., editor, *Advances in Neural Information Processing Systems 3*, pages 875–882. Morgan Kaufmann, 1991.

[111] N. Weir. Applications of maxmimum entropy techniques to HST data. In *Proceedings of the ESO/ST–ECF Data Analysis Workshop, April 1991*, 1991.

[112] D. Willshaw and P. Dayan. Optimal plasticity from matrix memories: what goes up must come down. *Neural Computation*, 2(1):85–93, 1990.