Bibliography

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

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

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

[4] J.R.P. Angel, P. Wizinowich, M. Lloyd-Hart, and D. Sandler (1990). Adaptive optics for array telescopes using neural–network techniques, *Nature* **348**, 221–224.

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

[6] S. Becker and Y. Le Cun (1988). Improving the convergence of back–propagation learning with second order methods, in *Proc. of the connectionist models Summer school,* Ed. D.S. Touretzky *et. al.*, 29, Morgan Kaufmann.

[7] J. Berger (1985). *Statistical decision theory and Bayesian analysis,* Springer.

[8] C.M. Bishop (1991). Exact calculation of the Hessian matrix for the multilayer perceptron, submitted to *Neural Computation.*

[9] G.L. Bretthorst (1990). Bayesian Analysis. I. Parameter Estimation Using Quadrature NMR Models. II. Signal Detection and Model Selection. III. Applications to NMR., *J. Magnetic Resonance* **88** 3, 533–595.

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

[11] M.K. Charter (1991). Quantifying drug absorption, in [20], 245–252.

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

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

[14] J.S. Denker and Y. Le Cun (1991). Transforming neural-net output levels to probability distributions, in *Advances in neural information processing systems 3,* ed. R.P. Lippmann *et. al.*, 853–859, Morgan Kaufmann.

[15] R. Duda and P. Hart (1973). *Pattern Classification and Scene Analysis,* Wiley.

[16] M.A. El–Gamal (1991). The role of priors in active Bayesian learning in the sequential statistical decision framework, in [20], 33–38.

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

[18] V.V. Fedorov (1972). *Theory of optimal experiments,* Academic press.

[19] K. Fukunaga (1972). *Introduction to statistical pattern recognition,* Academic press.

[20] W.T. Grandy, Jr. and L.H. Schick, eds. (1991). *Maximum Entropy and Bayesian Methods, Laramie 1990,* Kluwer.

[21] S.F. Gull (1988). Bayesian inductive inference and maximum entropy, in *Maximum Entropy and Bayesian Methods in science and engineering, vol. 1: Foundations*, G.J. Erickson and C.R. Smith, eds., Kluwer.

[22] S.F. Gull (1989). Developments in Maximum entropy data analysis, in [61], 53–71.

[23] S.F. Gull (1989). Bayesian data analysis: straight–line fitting, in [61], 511–518.

[24] S.F. Gull and J. Skilling (1991). *Quantified Maximum Entropy.* `MemSys5` *User's manual,* M.E.D.C., 33 North End, Royston, SG8 6NR, England.

[25] R. Hanson, J. Stutz and P. Cheeseman (1991). Bayesian classification theory, *NASA Ames* TR FIA–90-12-7-01.

[26] D. Haussler, M. Kearns and R. Schapire (1991). Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension, preprint.

[27] G.E. Hinton and T.J. Sejnowski (1986). Learning and relearning in Boltzmann machines, in *Parallel Distributed Processing,* Rumelhart *et. al.*, MIT Press.

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

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

[30] E.T. Jaynes (1986). Bayesian methods: general background, in *Maximum Entropy and Bayesian Methods in applied statistics,* ed. J.H. Justice, C.U.P..

[31] H. Jeffreys (1939). *Theory of Probability,* Oxford Univ. Press.

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

[33] R.L. Kashyap (1977). A Bayesian comparison of different classes of dynamic models using empirical data, *IEEE Transactions on Automatic Control* **AC-22** 5, 715–727.

[34] Y. Le Cun, J.S. Denker and S.S. Solla (1990). Optimal Brain Damage, in *Advances in neural information processing systems 2,* ed. David S. Touretzky, 598–605, Morgan Kaufmann.

[35] W.T. Lee and M.F. Tenorio (1991). On Optimal Adaptive Classifier Design Criterion — How many hidden units are necessary for an optimal neural network classifier?, *Purdue University* TR-EE-91-5.

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

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

[38] T.J. Loredo (1989). From Laplace to supernova SN 1987A: Bayesian inference in astrophysics, in *Maximum Entropy and Bayesian Methods,* ed. P. Fougere, Kluwer.

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

[40] D.J.C. MacKay (1991). Bayesian interpolation, to appear in *Neural Computation*; Chapter 2 of this dissertation.

[41] D.J.C. MacKay (1991). A practical Bayesian framework for backprop networks, to appear in *Neural Computation*; Chapter 3 of this dissertation.

[42] D.J.C. MacKay (1991). Information–based objective functions for active data selection, to appear in *Neural Computation*; Chapter 4 of this dissertation.

[43] D.J.C. MacKay (1991). The evidence framework applied to classification networks, submitted to *Neural Computation*; Chapter 5 of this dissertation.

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

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

[46] R.M. Neal (1991). Bayesian mixture modeling by Monte Carlo simulation, preprint.

[47] S.J. Nowlan (1991). Soft competitive adaptation: neural network learning algorithms based on fitting statistical mixtures, *Carnegie Mellon University* Doctoral thesis CS–91–126.

[48] S.J. Nowlan and G.E. Hinton (1991). Soft weight sharing, preprint.

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

[50] J.D. Patrick and C.S. Wallace (1982). Stone circle geometries: an information theory approach, in *Archaeoastronomy in the Old World,* D.C. Heggie, editor, Cambridge Univ. Press.

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

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

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

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

[55] J. Rissanen (1978). Modeling by shortest data description, *Automatica* **14**, 465–471.

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

[57] D.E. Rumelhart (1987). Cited in [32].

[58] G. Schwarz (1978). Estimating the dimension of a model, *Ann. Stat.* **6** 2, 461–464.

[59] H.S. Seung, H. Sompolinsky and N. Tishby (1991). Statistical mechanics of learning from examples, preprint.

[60] S. Sibisi (1991). Bayesian interpolation, in [20], 349–355.

[61] J. Skilling, editor (1989). *Maximum Entropy and Bayesian Methods, Cambridge 1988,* Kluwer.

[62] J. Skilling (1989). The eigenvalues of mega–dimensional matrices, in [61], 455–466.

[63] J. Skilling (1991). On parameter estimation and quantified MaxEnt, in [20], 267–273.

[64] J. Skilling, D.R.T. Robinson, and S.F. Gull (1991). Probabilistic displays, in [20], 365–368.

[65] J. Skilling (1991). Fundamentals of MaxEnt in data analysis, in *Maximum Entropy in action*, B. Buck and V.A. MacAulay, eds., Oxford, 19–40..

[66] J. Skilling (1992). Bayesian solution of ordinary differential equations, in *Maximum Entropy and Bayesian Methods, Seattle 1991,* G.J. Erickson and C.R. Smith, eds., Kluwer.

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

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

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

[70] R. Szeliski (1989). *Bayesian modeling of uncertainty in low level vision,* Kluwer.

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

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

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

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

[75] C. S. Wallace and P. R. Freeman (1987). Estimation and Inference by Compact Coding, *J. R. Statist. Soc. B* **49** 3, 240-265.

[76] N. Weir (1991). Applications of maximum entropy techniques to HST data, in Proceedings of the ESO/ST–ECF Data Analysis Workshop, April 1991.

[77] A.S. Weigend, D.E. Rumelhart and B.A. Huberman (1991). Generalization by weight–elimination with applications to forecasting, in *Advances in neural information processing systems 3,* ed. R.P. Lippmann *et. al.*, 875–882, Morgan Kaufmann.

[78] A. Zellner (1984). *Basic issues in econometrics,* Chicago.