

## References

- Adriaans, P., and C. Jacobs (2006). Using MDL for grammar induction. In *Proceedings of the Eighth International Colloquium on Grammatical Inference (ICGI-2006)*. To appear.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *Second International Symposium on Information Theory*, pp. 267–281.
- Allen, T. V., and R. Greiner (2000). Model selection criteria for learning belief nets: An empirical comparison. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML-97)*.
- Allen, T. V., O. Madani, and R. Greiner (2003). Comparing model selection criteria for belief networks. Submitted.
- Anthony, M., and N. Biggs (1992). *Computational Learning Theory*. Cambridge, UK: Cambridge University Press.
- Bain, L., and M. Engelhardt (1989). *Introduction to Probability and Mathematical Statistics*. Boston: PWS-Kent.
- Balasubramanian, V. (1997). Statistical inference, Occam’s razor, and statistical mechanics on the space of probability distributions. *Neural Computation* 9, 349–368.
- Balasubramanian, V. (2005). MDL, Bayesian inference and the geometry of the space of probability distributions. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.
- Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. Chichester, UK: Wiley.

- Barron, A. (1985). *Logically Smooth Density Estimation*. Ph. D. thesis, Department of Electrical Engineering, Stanford University, Stanford, CA.
- Barron, A. (1986). Discussion on Diaconis and Freedman: the consistency of Bayes estimates. *Annals of Statistics* 14, 26–30.
- Barron, A. (1990). Complexity regularization with application to artificial neural networks. In G. Roussas (Ed.), *Nonparametric Functional Estimation and Related Topics*, pp. 561–576. Dordrecht, the Netherlands: Kluwer Academic Publishers.
- Barron, A. (1998). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In A. D. J.M. Bernardo, J.O. Berger and A. Smith (Eds.), *Bayesian Statistics*, volume 6, pp. 27–52. Oxford: Oxford University Press.
- Barron, A., and T. Cover (1991). Minimum complexity density estimation. *IEEE Transactions on Information Theory* 37(4), 1034–1054.
- Barron, A., and N. Hengartner (1998). Information theory and superefficiency. *Annals of Statistics* 26(5), 1800–1825.
- Barron, A., J. Rissanen, and B. Yu (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory* 44(6), 2743–2760. Special Commemorative Issue: Information Theory: 1948-1998.
- Barron, A., and C. Sheu (1991). Approximation of density functions by sequences of exponential families. *Annals of Statistics* 19(3), 1347–1369.
- Barron, A., Y. Yang, and B. Yu (1994). Asymptotically optimal function estimation by minimum complexity criteria. In *Proceedings of the 1994 International Symposium on Information Theory*, pp. 38. Trondheim, Norway.
- Bartlett, P., S. Boucheron, and G. Lugosi (2001). Model selection and error estimation. *Machine Learning* 48, 85–113.
- Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*, revised and expanded 2nd edition. Springer Series in Statistics. New York: Springer-Verlag.
- Berger, J., and L. Pericchi (2001). Objective Bayesian methods for model selection: introduction and comparison. In P. Lahiri (Ed.), *Model Selection*, volume 38 of *IMS Lecture Notes – Monograph Series*, pp. 135–207. Beachwood, Ohio: Institute of Mathematical Statistics.

- Bernardo, J., and A. Smith (1994). *Bayesian Theory*. Chichester: Wiley.
- Blackwell, D., and L. Dubins (1962). Merging of opinions with increasing information. *Annals of Mathematical Statistics* 33, 882–886.
- Blum, A., and J. Langford (2003). PAC-MDL bounds. In *Proceedings of the Sixteenth Conference on Learning Theory (COLT' 03)*, pp. 344–357.
- Blumer, A., A. Ehrenfeucht, D. Haussler, and M. Warmuth (1987). Occam's razor. *Information Processing Letters* 24, 377–380.
- Boucheron, S., O. Bousquet, and G. Lugosi (2005). Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics* 9, 323–375.
- Breiman, L. (2001). Statistical modeling: the two cultures (with discussion). *Statistical Science* 16(3), 199–215.
- Brown, L. (1986). *Fundamentals of Statistical Exponential Families, with Applications in Statistical Decision Theory*. Hayward, CA: Institute of Mathematical Statistics.
- Burnham, K., and D. Anderson (2002). *Model Selection and Multimodel Inference*. New York: Springer-Verlag.
- Casella, G., and R. Berger. *Statistical Inference*. Belmont, CA: Wadsworth.
- Cesa-Bianchi, N., Y. Freund, D. Helmbold, D. Haussler, R. Schapire, and M. Warmuth (1997). How to use expert advice. *Journal of the ACM* 44(3), 427–485.
- Cesa-Bianchi, N., and G. Lugosi (2006). *Prediction, Learning and Games*. Cambridge, UK: Cambridge University Press.
- Chaitin, G. (1966). On the length of programs for computing finite binary sequences. *Journal of the ACM* 13, 547–569.
- Chaitin, G. (1969). On the length of programs for computing finite binary sequences: Statistical considerations. *Journal of the ACM* 16, 145–159.
- Chernoff, H. (1952). A measure of asymptotic efficiency of test of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics* 23, 493–507.
- Cilibrasi, R., and P. Vitányi (2005). Clustering by compression. *IEEE Transactions on Information Theory* 51(4), 1523–1545.
- Clarke, B. (1997). Online forecasting proposal. Technical report, University of Dortmund. Sonderforschungsbereich 475.

- Clarke, B. (2003). Combining model selection procedures for online prediction. *Sankhyā: The Indian Journal of Statistics, Series A* 63, 229–249.
- Clarke, B. (2004). Comparing Bayes and non-Bayes model averaging when model approximation error cannot be ignored. *Journal of Machine Learning Research* 4(4), 683–712.
- Clarke, B., and A. Barron (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory IT-36*(3), 453–471.
- Clarke, B., and A. Barron (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference* 41, 37–60.
- Clarke, B., and A. Dawid (1999). Online prediction with experts under a log-scoring rule. Unpublished manuscript.
- Comley, J. W., and D. L. Dowe (2005). Minimum message length and generalized Bayesian nets with asymmetric languages. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.
- Conway, J., and N. Sloane (1993). *Sphere Packings, Lattices and Groups*. New York: Springer-Verlag.
- Cover, T., and J. Thomas (1991). *Elements of Information Theory*. New York: Wiley-Interscience.
- Cramér, H. (1938). Sur un nouveau théorème-limite de la théorie des probabilités. *Actualités Scientifiques et Industrielles* 736, 5–23.
- Csiszár, I. (1984). Sanov property, generalized  $I$ -projection and a conditional limit theorem. *Annals of Probability* 12(3), 768–793.
- Csiszár, I., and P. Shields (2000). The consistency of the BIC Markov order estimator. *Annals of Statistics* 28, 1601–1619.
- Davies, P., and A. Kovac (2001). Modality, runs, strings and multiresolution (with discussion). *Annals of Statistics* 29, 1–65.
- Dawid, A. (1984). Present position and potential developments: Some personal views, statistical theory, the prequential approach. *Journal of the Royal Statistical Society, Series A* 147(2), 278–292.
- Dawid, A. (1992). Prequential analysis, stochastic complexity and Bayesian inference. In J. Bernardo, J. Berger, A. Dawid, and A. Smith

- (Eds.), *Bayesian Statistics*, volume 4, pp. 109–125. Oxford: Oxford University Press.
- Dawid, A. (1997). Prequential analysis. In S. Kotz, C. Read, and D. Banks (Eds.), *Encyclopedia of Statistical Sciences*, volume 1 (Update), pp. 464–470. New York: Wiley-Interscience.
- Dawid, A. P., and V. G. Vovk (1999). Prequential probability: Principles and properties. *Bernoulli* 5, 125–162.
- De Finetti, B. (1937). La prevision: ses lois logiques, ses sources subjectives. *Annales Institut H. Poincaré* 7, 1–68.
- De Finetti, B. (1974). *Theory of Probability. A Critical Introductory Treatment*. London: Wiley.
- De Luna, X., and K. Skouras (2003). Choosing a model selection strategy. *Scandinavian Journal of Statistics* 30, 113–128.
- De Rooij, S., and P. D. Grünwald (2006). An empirical study of MDL model selection with infinite parametric complexity. *Journal of Mathematical Psychology* 50(2), 180–192.
- Devroye, L., L. Györfi, and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag.
- Diaconis, P. (2003). The problem of thinking too much. *Bulletin of the American Academy of Arts and Sciences* 16(3), 26–38.
- Diaconis, P., and D. Freedman (1986). On the consistency of Bayes estimates. *The Annals of Statistics* 14(1), 1–26.
- Domingos, P. (1999). The role of Occam’s razor in knowledge discovery. *Data Mining and Knowledge Discovery* 3(4), 409–425.
- Doob, J. (1949). Application of the theory of martingales. In *Le Calcul de Probabilités et ses Applications. Colloques Internationaux du Centre National de la Recherche Scientifique*, pp. 23–27.
- Drmotá, M., and W. Szpankowski (2004). Precise minimax redundancy and regret. *IEEE Transactions on Information Theory* 50, 2686–2707.
- Duda, R., P. Hart, and D. Stork (2000). *Pattern Classification*. New York: Wiley.
- Elias, P. (1975). Universal codeword sets and representation of the integers. *IEEE Transactions on Information Theory* 21(2), 194–203.
- Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics* 75, 643–649.

- Feder, M. (1986). Maximum entropy as a special case of the minimum description length criterion. *IEEE Transactions on Information Theory* 32(6), 847–849.
- Feller, W. (1968a). *An Introduction to Probability Theory and Its Applications*, 3rd edition, volume 1. New York: Wiley.
- Feller, W. (1968b). *An Introduction to Probability Theory and Its Applications*, 3rd edition, volume 2. New York: Wiley.
- Ferguson, T. (1967). *Mathematical Statistics – a decision-theoretic approach*. San Diego: Academic Press.
- Figueiredo, M., J. Leitão, and A.K.Jain (2000). Unsupervised contour representation and estimation using b-splines and a minimum description length criterion. *IEEE Transactions on Image Processing* 9(6), 1075–1087.
- Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A* 222, 309–368.
- Floyd, S., and M. Warmuth (1995). Sample compression, learnability and the Vapnik-Chervonenkis dimension. *Machine Learning* 21, 269–304.
- Forster, M. (2001). The new science of simplicity. In A. Zellner, H. Keuzenkamp, and M. McAleer (Eds.), *Simplicity, Inference and Modelling*, pp. 83–117. Cambridge: Cambridge University Press.
- Foster, D., and R. Stine (1999). Local asymptotic coding and the minimum description length. *IEEE Transactions on Information Theory* 45, 1289–1293.
- Foster, D., and R. Stine (2001). The competitive complexity ratio. In *Proceedings of the 2001 Conference on Information Sciences and Systems*. WP8 1-6.
- Foster, D. P., and R. A. Stine (2005). The contribution of parameters to stochastic complexity. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.
- Freund, Y. (1996). Predicting a binary sequence almost as well as the optimal biased coin. In *Proceedings of the Ninth Annual ACM Conference on Computational Learning Theory (COLT' 96)*, pp. 89–98.
- Friedman, N., D. Geiger, and M. Goldszmidt (1997). Bayesian network classifiers. *Machine Learning* 29, 131–163.

- Gács, P., J. Tromp, and P. Vitányi (2001). Algorithmic statistics. *IEEE Transactions on Information Theory* 47(6), 2464–2479.
- Gao, Q., and M. Li (1989). An application of minimum description length principle to online recognition of handprinted alphanumerals. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI-89)*, pp. 843–848.
- Gauss, C. (1957). *Gauss's Work on the Theory of Least Squares (1803-1826)*. Princeton, NJ: Princeton University Press. Translated by H.F. Trotter.
- Gelman, A., B. Carlin, H. Stern, and D. Rubin (2003). *Bayesian Data Analysis*. Boca Raton, FL: CRC Press.
- George, E., and D. Foster (2000). Calibration and empirical Bayes variable selection. *Biometrika* 84(4), 731–747.
- Gerencsér, L. (1987). Order estimation of stationary Gaussian ARMA processes using Rissanen's complexity. Technical report, Computer and Automation Institute of the Hungarian Academy of Sciences.
- Gerencsér, L. (1994). On Rissanen's predictive stochastic complexity for stationary ARMA processes. *Journal of Statistical Planning and Inference* 41, 303–325.
- Gibbs, A., and F. Su (2002). On choosing and bounding probability metrics. *International Statistical Review* 70(3), 419–435.
- Goodman, N. (1955). *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.
- Grünwald, P. D. (1996). A minimum description length approach to grammar inference. In S. Wermter, E. Riloff, and G. Scheler (Eds.), *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, Number 1040 in Lecture Notes in Artificial Intelligence, pp. 203–216. New York: Springer-Verlag.
- Grünwald, P. D. (1998). *The Minimum Description Length Principle and Reasoning under Uncertainty*. Ph. D. thesis, University of Amsterdam, the Netherlands. Available as ILLC Dissertation Series 1998-03.
- Grünwald, P. D. (1999). Viewing all models as “probabilistic”. In *Proceedings of the Twelfth ACM Conference on Computational Learning Theory (COLT' 99)*, pp. 171–182.
- Grünwald, P. D. (2000). Maximum entropy and the glasses you are looking through. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI 2000)*, pp. 238–246.

- Grünwald, P. D. (2001). Strong entropy concentration, game theory and algorithmic randomness. In *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory (COLT' 01)*, pp. 320–336.
- Grünwald, P. D. (2005). A tutorial introduction to the minimum description principle. In P. D. Grünwald, I. Myung, and M. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*, pp. 3–79. Cambridge, MA: MIT Press.
- Grünwald, P. D. (2007). Prediction is coding. Manuscript in preparation.
- Grünwald, P. D., and S. de Rooij (2005). Asymptotic log-loss of prequential maximum likelihood codes. In *Proceedings of the Eighteenth Annual Conference on Computational Learning Theory (COLT 2005)*, pp. 652–667.
- Grünwald, P. D., and J. Y. Halpern (2003). Updating probabilities. *Journal of Artificial Intelligence Research* 19, 243–278.
- Grünwald, P. D., and J. Langford (2004). Suboptimality of MDL and Bayes in classification under misspecification. In *Proceedings of the Seventeenth Conference on Learning Theory (COLT' 04)*.
- Grünwald, P. D., and J. Langford (2007). Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*. To appear.
- Grünwald, P. D., I. J. Myung, and M. A. Pitt (Eds.) (2005). *Advances in Minimum Description Length: Theory and Applications*. MIT Press.
- Hall, P., and E. Hannan (1988). On stochastic complexity and nonparametric density estimation. *Biometrika* 75, 705–714.
- Halpern, J. (2003). *Reasoning about Uncertainty*. Cambridge, MA: MIT Press.
- Hannan, E. (1980). The estimation of the order of an ARMA process. *Annals of Statistics* 8, 1071–1081.
- Hannan, E., A. McDougall, and D. Poskitt (1989). Recursive estimation of autoregressions. *Journal of the Royal Statistical Society, Series B* 51, 217–233.
- Hannan, E., and J. Rissanen (1982). Recursive estimation of mixed autoregressive-moving average order. *Biometrika* 69, 81–94.
- Hansen, M., and B. Yu (2000). Wavelet thresholding via MDL for natural images. *IEEE Transactions on Information Theory* 46, 1778–1788.



- Hansen, M., and B. Yu (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association* 96(454), 746–774.
- Hansen, M., and B. Yu (2002). Minimum description length model selection criteria for generalized linear models. In *Science and Statistics: Festschrift for Terry Speed*, volume 40 of *IMS Lecture Notes – Monograph Series*. Hayward, CA: Institute for Mathematical Statistics.
- Hanson, A. J., and P. C.-W. Fu (2005). Applications of MDL to selected families of models. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.
- Harremoës, P. (2004). The weak information projection. In *Proceedings of the 2004 International Symposium on Information Theory (ISIT 2004)*, pp. 28.
- Harremoës, P. (2006). Interpretations of Rényi entropies and divergences. *Physica A* 365(1), 57–62.
- Harremoës, P., and F. Topsøe (2001). Maximum entropy fundamentals. *Entropy* 3, 191–226. Available at <http://www.mdpi.org/entropy/>.
- Hartigan, J. (1983). *Bayes Theory*. New York: Springer-Verlag.
- Haussler, D. (1997). A general minimax result for relative entropy. *IEEE Transactions on Information Theory* 43(4), 1276–1280.
- Haussler, D., and M. Opper (1997). Mutual information, metric entropy, and cumulative relative entropy risk. *Annals of Statistics* 25(6), 2451–2492.
- Helmbold, D., and M. Warmuth (1995). On weak learning. *Journal of Computer and System Sciences* 50, 551–573.
- Hemerly, E., and M. Davis (1989a). Recursive order estimation of stochastic control systems. *Mathematical Systems Theory* 22, 323–346.
- Hemerly, E., and M. Davis (1989b). Strong consistency of the PLS criterion for order determination of autoregressive processes. *Annals of Statistics* 17(2), 941–946.
- Herbrich, R. (2002). *Learning Kernel Classifiers*. Cambridge, MA: MIT Press.
- Herbrich, R., and R. C. Williamson (2002). Algorithmic luckiness. *Journal of Machine Learning Research* 3, 175–212.

- Hertz, J., A. Krogh, and R. Palmer (1991). *Introduction to the theory of neural computation*. Lecture Notes of the Santa Fe Institute. Boston: Addison-Wesley.
- Hjorth, U. (1982). Model selection and forward validation. *Scandinavian Journal of Statistics* 9, 95–105.
- Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* 58, 13–30.
- Hoerl, A., and R. Kennard (1970). Ridge regression: Biased estimation of non-orthogonal components. *Technometrics* 12, 55–67.
- Hutter, M. (2003). Optimality of universal Bayesian sequence prediction for general loss and alphabet. *Journal of Machine Learning Research* 4, 971–1000.
- Hutter, M. (2004). *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Berlin: Springer-Verlag.
- Hutter, M. (2006). On the foundations of universal sequence prediction. In *Proceedings of the Third Annual Conference on Theory and Applications of Models of Computation (TAMC 2006)*, pp. 408–420.
- Jaynes, E. (1957). Information theory and statistical mechanics. *Physical Review* 106(4), 620–630.
- Jaynes, E. (2003). *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge University Press. Edited by G. Larry Bretthorst.
- Jeffereys, W., and J. Berger (1992). Ockham's razor and Bayesian analysis. *American Scientist* 80, 64–72.
- Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Statistical Society (London) Series A* 186, 453–461.
- Jeffreys, H. (1961). *Theory of Probability*, 3rd edition. London: Oxford University Press.
- Jornsten, R., and B. Yu (2003). Simultaneous gene clustering and subset selection for classification via mdl. *Bioinformatics* 19(9), 1100–1109.
- Kakade, S., M. Seeger, and D. Foster (2006). Worst-case bounds for Gaussian process models. In *Proceedings of the 2005 Neural Information Processing Systems Conference (NIPS 2005)*.
- Kallenberg, O. (2002). *Foundations of Modern Probability*, 2nd edition. New York: Springer-Verlag.

- Kalnishkan, Y., and M. Vyugin (2002). Mixability and the existence of weak complexities. In *Proceedings of the Fifteenth Conference on Computational Learning Theory (COLT' 02)*, pp. 105–120.
- Kapur, J. N., and H. K. Kesavan (1992). *Entropy Optimization Principles with Applications*. San Diego: Academic Press.
- Kass, R., and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association* 90(430), 773–795.
- Kass, R., and P. Voss (1997). *Geometrical Foundations of Asymptotic Inference*. New York: Wiley-Interscience.
- Kass, R., and L. Wasserman (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association* 91, 1343–1370.
- Kearns, M., Y. Mansour, A. Ng, and D. Ron (1997). An experimental and theoretical comparison of model selection methods. *Machine Learning* 27, 7–50.
- Kelly, J. (1956). A new interpretation of information rate. *Bell System Technical Journal*, 917–926.
- Kolmogorov, A. (1941). Interpolation und Extrapolation von stationären zufälligen Folgen. *Izvestiia Akademii Nauk SSSR* 5, 3–14.
- Kolmogorov, A. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission* 1(1), 1–7.
- Kolmogorov, A. (1974a). Talk at the Information Theory Symposium in Tallinn, Estonia, 1974, according to P. Gács and T. Cover who attended it.
- Kolmogorov, A. (1974b). Complexity of algorithms and objective definition of randomness. A talk at Moscow Mathematical Society meeting, April 16th, 1974. A 4-line abstract is available in *Uspekhi Matematicheskikh Nauk* 29:4(1974), 155 (in Russian).
- Kontkanen, P., W. Buntine, P. Myllymäki, J. Rissanen, and H. Tirri (2003). Efficient computation of stochastic complexity. In C. Bishop and B. Frey (Eds.), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics (AISTATS 2003)*, pp. 181–188.
- Kontkanen, P., and P. Myllymäki (2005a). Analyzing the stochastic complexity via tree polynomials. Unpublished manuscript.
- Kontkanen, P., and P. Myllymäki (2005b). A fast normalized maximum likelihood algorithm for multinomial data. In *Proceedings of the Nine-*

- teenth International Joint Conference on Artificial Intelligence (IJCAI-05), pp. 1613–1616.
- Kontkanen, P., P. Myllymäki, W. Buntine, J. Rissanen, and H. Tirri (2005). An MDL framework for data clustering. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.
- Kontkanen, P., P. Myllymäki, T. Silander, and H. Tirri (1999). On supervised selection of Bayesian networks. In K. Laskey and H. Prade (Eds.), *Proceedings of the Fifteenth International Conference on Uncertainty in Artificial Intelligence (UAI'99)*.
- Kontkanen, P., P. Myllymäki, T. Silander, H. Tirri, and P. D. Grünwald (1998). Bayesian and information-theoretic priors for Bayesian network parameters. In C. Nedellec and C. Rouveirol (Eds.), *Machine Learning: ECML-98, Proceedings of the Tenth European Conference*, volume 1398 of *Lecture Notes in Artificial Intelligence*, pp. 89–94.
- Kraft, L. (1949). A device for quantizing, grouping and coding amplitude modulated pulses. Master's thesis, Department of Electrical Engineering, MIT, Cambridge, MA.
- Krichevsky, R., and V. Trofimov (1981). The performance of universal encoding. *IEEE Transactions on Information Theory* 27, 199–207.
- Lai, T., and C. Lee (1997). Information and prediction criteria for model selection in stochastic regression and ARMA models. *Statistica Sinica* 7, 285–309.
- Lanterman, A. (2001). Schwarz, Wallace, and Rissanen: Intertwining themes in theories of model order estimation. *International Statistical Review* 69(2), 185–212.
- Lanterman, A. D. (2005). Hypothesis testing for Poisson versus geometric distributions using stochastic complexity. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.
- Lee, M. (2002a). Generating additive clustering models with minimal stochastic complexity. *Journal of Classification* 19(1), 69–85.
- Lee, M. (2002b). A simple method for generating additive clustering models with limited complexity. *Machine Learning* 49, 39–58.
- Lee, P. (1997). *Bayesian Statistics — An Introduction*. London and Oxford: Arnold & Oxford University Press.

- Lee, T. (2000). Regression spline smoothing using the minimum description length principle. *Statistics and Probability Letters* 48(71–82).
- Lee, T. (2002c). Automatic smoothing for discontinuous regression functions. *Statistica Sinica* 12, 823–842.
- Levenstein, V. (1968). On the redundancy and delay of separable codes for the natural numbers. *Problems of Cybernetics* 20, 173–179.
- Li, J. (1999). *Estimation of Mixture Models*. Ph. D. thesis, Yale University, New Haven, CT.
- Li, J., and A. Barron (2000). Mixture density estimation. In S. Solla, T. Leen, and K.-R. Müller (Eds.), *Advances in Neural Information Processing Systems*, volume 12, pp. 279–285.
- Li, K. (1987). Asymptotic optimality of  $c_p$ ,  $c_l$ , cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics* 15, 958–975.
- Li, L., and B. Yu (2000). Iterated logarithmic expansions of the pathwise code lengths for exponential families. *IEEE Transactions on Information Theory* 46(7), 2683–2689.
- Li, M., and P. Vitányi (1997). *An Introduction to Kolmogorov Complexity and Its Applications*, revised and expanded 2nd edition. New York: Springer-Verlag.
- Liang, F., and A. Barron (2005). Exact minimax predictive density estimation and MDL. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.
- Liang, F., and A. R. Barron (2002). Exact minimax strategies for predictive density estimation, data compression, and model selection. In *Proceedings of the 2002 IEEE International Symposium on Information Theory (ISIT 2002)*.
- Liang, F., and A. R. Barron (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Transactions on Information Theory* 50, 2708–2726.
- Lindley, D., and A. Smith (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B* 34, 1–41. With discussion.
- Littlestone, N., and M. Warmuth (1994). The weighted majority algorithm. *Information and Computation* 108(2), 212–261.

- Liu, J., and P. Moulin (1998). A new complexity prior for multiresolution image denoising. In *Proceedings of IEEE Workshop on Time-Frequency Time-Scale Analysis*, pp. 637–640.
- Lutwak, E., D. Yang, and G. Zhang (2005). Cramér-Rao and moment-entropy inequalities for Rényi entropy and generalized Fisher information. *IEEE Transactions on Information Theory* 51, 473–478.
- MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge University Press.
- McAllester, D. (1998). Some PAC-Bayesian theorems. In *Proceedings of the Eleventh ACM Conference on Computational Learning Theory (COLT' 98)*, pp. 230–234.
- McAllester, D. (1999). PAC-Bayesian model averaging. In *Proceedings of the Twelfth ACM Conference on Computational Learning Theory (COLT' 99)*, pp. 164–171.
- McAllester, D. (2003). PAC-Bayesian stochastic model selection. *Machine Learning* 51(1), 5–21.
- Mehta, M., J. Rissanen, and R. Agrawal (1995). MDL-based decision tree pruning. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD '95)*, pp. 216–221.
- Meir, R., and N. Merhav (1995). On the stochastic complexity of learning realizable and unrealizable rules. *Machine Learning* 19, 241–261.
- Merhav, N., and M. Feder (1998). Universal prediction. *IEEE Transactions on Information Theory* IT-44(6), 2124–2147. Special Commemorative Issue: Information Theory: 1948-1998.
- Michalski, R., J. Carbonell, and T. Mitchell (1983). *Machine Learning, An Artificial Intelligence Approach*. San Francisco: Morgan Kaufmann.
- Michie, D., D. Spiegelhalter, and C. Taylor (Eds.) (1994). *Machine Learning, Neural and Statistical Classification*. London: Ellis Horwood.
- Modha, D. S., and E. Masry (1998). Prequential and cross-validated regression estimation. *Machine Learning* 33(1), 5–39.
- Myung, I., V. Balasubramanian, and M. Pitt (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences USA* 97, 11170–11175.
- Myung, I. J., M. A. Pitt, S. Zhang, and V. Balasubramanian (2000). The use of MDL to select among computational models of cognition. In

- Advances in Neural Information Processing Systems*, volume 13, pp. 38–44. Cambridge, MA: MIT Press.
- Nannen, V. (2003). The paradox of overfitting. Master's thesis, University of Groningen, Groningen, the Netherlands.
- Navarro, D. (2004). A note on the applied use of MDL approximations. *Neural Computation* 16, 1763–1768.
- Ndili, U., R. Nowak, and M. Figueiredo (2001). Coding-theoretic approach to image segmentation. In *Proceedings of the 2001 IEEE International Conference on Image Processing - ICIP'2001*.
- Neal, R. (1996). *Bayesian learning for neural networks*. New York: Springer-Verlag.
- Nowak, R., and M. Figueiredo (2000). Unsupervised segmentation of Poisson data. In *Proceedings of the International Conference on Pattern Recognition - ICPR'2000*, volume 3, pp. 159–162.
- Osborne, M. (1999). MDL-based DCG induction for NP identification. In *Proceedings of the Third Conference on Computational Natural Language Learning (CoNLL '99)*, pp. 61–68.
- Pednault, E. (2003). Personal communication, June 2003.
- Poland, J., and M. Hutter (2005). Asymptotics of discrete MDL for online prediction. *IEEE Transactions on Information Theory* 51(11), 3780–3795.
- Poland, J., and M. Hutter (2006). MDL convergence speed for Bernoulli sequences. *Statistics and Computing* 16, 161–175.
- Qian, G., G. Gabor, and R. Gupta (1996). Generalised linear model selection by the predictive least quasi-deviance criterion. *Biometrika* 83, 41–54.
- Qian, G., and H. Künsch (1998). Some notes on Rissanen's stochastic complexity. *IEEE Transactions on Information Theory* 44(2), 782–786.
- Quinlan, J., and R. Rivest (1989). Inferring decision trees using the minimum description length principle. *Information and Computation* 80, 227–248.
- Rasmussen, C., and Z. Ghahramani (2000). Occam's razor. In *Advances in Neural Information Processing Systems*, volume 13, pp. 294–300.
- Rasmussen, C., and C. Williams (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

- Rényi, A. (1960). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 547–561.
- Rice, J. (1995). *Mathematical Statistics and Data Analysis*. Duxbury Press.
- Ripley, B. (1996). *Pattern Recognition and Neural Networks*. Cambridge, UK: Cambridge University Press.
- Rissanen, J. (1978). Modeling by the shortest data description. *Automatica* 14, 465–471.
- Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics* 11, 416–431.
- Rissanen, J. (1984). Universal coding, information, prediction and estimation. *IEEE Transactions on Information Theory* 30, 629–636.
- Rissanen, J. (1986a). Order estimation by accumulated prediction errors. In J. Gani and M. B. Priestley (Eds.), *Essays in Time Series and Allied Processes*, pp. 55–61. Sheffield, UK: Applied Probability Trust.
- Rissanen, J. (1986b). A predictive least squares principle. *IMA Journal of Mathematical Control and Information* 3, 211–222.
- Rissanen, J. (1986c). Stochastic complexity and modeling. *Annals of Statistics* 14, 1080–1100.
- Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society, Series B* 49, 223–239. Discussion: 252–265.
- Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. Hackensack, NJ: World Scientific.
- Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory* 42(1), 40–47.
- Rissanen, J. (1999). Hypothesis selection and testing by the MDL principle. *Computer Journal* 42(4), 260–269.
- Rissanen, J. (2000). MDL denoising. *IEEE Transactions on Information Theory* 46(7), 2537–2543.
- Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory* 47(5), 1712–1717.
- Rissanen, J. (2007). *Information and Complexity in Statistical Modeling*. New York: Springer-Verlag.



- Rissanen, J., and E. Ristad (1994). Language acquisition in the MDL framework. In E. Ristad (Ed.), *Language Computations*. Philadelphia: American Mathematical Society.
- Rissanen, J., T. Speed, and B. Yu (1992). Density estimation by stochastic complexity. *IEEE Transactions on Information Theory* 38(2), 315–323.
- Rissanen, J., and I. Tabus (2005). Kolmogorov’s structure function in MDL theory and lossy data compression. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.
- Rissanen, J., and B. Yu (1995). MDL learning. In D. Kueker and C. Smith (Eds.), *Learning and Geometry: Computational Approaches, Progress in Computer Science and Applied Logic*, volume 14, pp. 3–19. Boston: Birkhäuser.
- Rockafellar, R. (1970). *Convex Analysis*. Princeton, NJ: Princeton University Press.
- Roos, T. (2004). MDL regression and denoising. Unpublished manuscript.
- Roos, T., P. Myllymäki, and H. Tirri (2005). On the behavior of MDL denoising. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS ’05)*, pp. 309–316.
- Roos, T., H. Wettig, P. Grünwald, P. Myllymäki, and H. Tirri (2005). On discriminative Bayesian network classifiers and logistic regression. *Machine Learning* 59(3), 267 – 296.
- Rosenfeld, R. (1996). A maximum entropy approach to adaptive statistical language modeling. *Computer, Speech and Language* 10, 187–228.
- Ross, S. (1998). *A First Course in Probability*. Upper Saddle River, NJ: Prentice-Hall.
- Rubin, H. (1987). A weak system of axioms for “rational” behavior and the nonseparability of utility from prior. *Statistical Decisions* 5, 47–58.
- Savage, L. (1954). *The Foundations of Statistics*. Dover Publications.
- Schölkopf, B., and A. J. Smola (2002). *Learning with Kernels*. Cambridge, MA: MIT Press.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6(2), 461–464.
- Seeger, M. (2004). Gaussian processes for machine learning. *International Journal of Neural Systems* 14(2), 69–106.

- Seidenfeld, T. (1986). Entropy and uncertainty. *Philosophy of Science* 53, 467–491.
- Shafer, G., and V. Vovk (2001). *Probability and Finance – It’s Only a Game!* New York: Wiley.
- Shaffer, C. (1993). Overfitting avoidance as bias. *Machine Learning* 10, 153–178.
- Shannon, C. (1948). The mathematical theory of communication. *Bell System Technical Journal* 27, 379–423, 623–656.
- Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association* 88, 486–494.
- Shao, J. (1997). An asymptotic theory for linear model selection (with discussion). *Statistica Sinica* 7, 221–242.
- Shawe-Taylor, J., P. Bartlett, R. Williamson, and M. Anthony (1998). Structural risk minimisation over data-dependent hierarchies. *IEEE Transactions on Information Theory* 44(5), 1926–1940.
- Shawe-Taylor, J., and N. Cristianini (2004). *Kernel Methods for Pattern Analysis*. Cambridge, UK: Cambridge University Press.
- Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike’s information criterion. *Biometrika* 63(1), 117–126.
- Shtarkov, Y. M. (1987). Universal sequential coding of single messages. *Problems of Information Transmission* 23(3), 3–17.
- Sober, E. (2004). The contest between parsimony and likelihood. *Systematic Biology* 4, 644–653.
- Solomonoff, R. (1964). A formal theory of inductive inference, part 1 and part 2. *Information and Control* 7, 1–22, 224–254.
- Solomonoff, R. (1978). Complexity-based induction systems: comparisons and convergence theorems. *IEEE Transactions on Information Theory* 24, 422–432.
- Speed, T., and B. Yu (1993). Model selection and prediction: Normal regression. *Annals of the Institute of Statistical Mathematics* 45(1), 35–54.
- Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* 64(4), 583–639.
- Starkie, B. (2001). Programming spoken dialogs using grammatical inference. In *Advances in Artificial Intelligence (AI 2001)*. Berlin: Springer.

- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B* 36(2), 111–147.
- Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B* 39, 44–47.
- Strang, G. (1988). *Linear Algebra and its Applications*, 3rd edition. Philadelphia: W.B. Saunders.
- Su, Y., I. J. Myung, and M. A. Pitt (2005). Minimum description length and cognitive modeling. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.
- Szpankowski, W. (1998). On asymptotics of certain recurrences arising in universal coding. *Problems of Information Transmission* 34(2), 142–146.
- Tabus, I., J. Rissanen, and J. Astola (2002). Normalized maximum likelihood models for Boolean regression with application to prediction and classification in genomics. In W. Zhang and I. Shmulevich (Eds.), *Computational and Statistical Approaches to Genomics*.
- Tabus, I., J. Rissanen, and J. Astola (2003). Classification and feature gene selection using the normalized maximum likelihood model for discrete regression. *Signal Processing* 83(4), 713–727. Special issue on Genomic Signal Processing.
- Takeuchi, J. (2000). On minimax regret with respect to families of stationary stochastic processes [in Japanese]. In *Proceedings IBIS 2000*, pp. 63–68.
- Takeuchi, J., and A. Barron (1997). Asymptotically minimax regret for exponential families. In *Proceedings SITA '97*, pp. 665–668.
- Takeuchi, J., and A. Barron (1998a). Asymptotically minimax regret by Bayes mixtures. In *Proceedings of the 1998 International Symposium on Information Theory (ISIT 98)*.
- Takeuchi, J., and A. R. Barron (1998b). Robustly minimax codes for universal data compression. In *Proceedings of the Twenty-First Symposium on Information Theory and Its Applications (SITA '98)*.
- Topsøe, F. (1979). Information-theoretical optimization techniques. *Kybernetika* 15(1), 8–27.

- Topsøe, F. (2007). Information theory at the service of science. In I. Csiszár, G. Katona, and G. Tardos (Eds.), *Entropy, Search, Complexity*, volume 16 of *Bolyai Society Mathematical Studies*. New York: Springer-Verlag.
- Townsend, P. (1975). The mind-body equation revisited. In C.-Y. Cheng (Ed.), *Psychological Problems in Philosophy*, pp. 200–218. Honolulu: University of Hawaii Press.
- Uffink, J. (1995). Can the maximum entropy principle be explained as a consistency requirement? *Studies in History and Philosophy of Modern Physics* 26B, 223–261.
- van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, UK: Cambridge University Press.
- Vapnik, V. (1982). *Estimation of Dependencies Based on Empirical Data*. Berlin: Springer-Verlag.
- Vapnik, V. (1998). *Statistical Learning Theory*. New York: Wiley.
- Vapnik, V., and A. Chervonenkis (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications* 16(2), 264–280.
- Vereshchagin, N., and P. Vitányi (2002). Kolmogorov’s structure functions with an application to the foundations of model selection. In *Proceedings Forty-Seventh IEEE Symposium on the Foundations of Computer Science (FOCS’02)*.
- Vereshchagin, N., and P. Vitányi (2004). Kolmogorov’s structure functions and model selection. *IEEE Transactions on Information Theory* 50(12), 3265–3290.
- Viswanathan, M., C. Wallace, D. Dowe, and K. Korb (1999). Finding cut-points in noisy binary sequences - A revised empirical evaluation. In *Proceedings of the Twelfth Australian Joint Conference on Artificial Intelligence*, volume 1747 of *Lecture Notes in Artificial Intelligence (LNAI)*, pp. 405–416.
- Vitányi, P. M. (2005). Algorithmic statistics and Kolmogorov’s structure function. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.
- von Neumann, J. (1928). Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100, 295–320.

- Vovk, V. (1990). Aggregating strategies. In *Proceedings of the Third Annual ACM Conference on Computational Learning Theory (COLT' 90)*, pp. 371–383.
- Vovk, V. (2001). Competitive on-line statistics. *International Statistical Review* 69, 213–248.
- Wagenmakers, E., P. D. Grünwald, and M. Steyvers (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology* 50(2), 149–166.
- Wallace, C. (2005). *Statistical and Inductive Inference by Minimum Message Length*. New York: Springer-Verlag.
- Wallace, C., and D. Boulton (1968). An information measure for classification. *Computer Journal* 11, 185–195.
- Wallace, C., and D. Boulton (1975). An invariant Bayes method for point estimation. *Classification Society Bulletin* 3(3), 11–34.
- Wallace, C., and P. Freeman (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society, Series B* 49, 240–251. Discussion: pages 252–265.
- Wallace, C., and J. Patrick (1993). Coding decision trees. *Machine Learning* 11, 7–22.
- Watanabe, S. (1999a). Algebraic analysis for non-regular learning machines. In *Advances in Neural Information Processing Systems*, pp. 356–363.
- Watanabe, S. (1999b). Algebraic analysis for singular statistical estimation. In *Tenth International Conference on Algorithmic Learning Theory (ALT'99)*, volume 1720 of *Lecture Notes in Computer Science*, pp. 39–50.
- Wax, M. (1988). Order selection for AR models by predictive least squares. *IEEE Transactions on Acoustics, Speech and Signal Processing* 36(4), 581–588.
- Webb, G. (1996). Further experimental evidence against the utility of Occam's razor. *Journal of Artificial Intelligence Research* 4, 397–417.
- Wei, C. (1992). On predictive least squares principles. *Annals of Statistics* 20(1), 1–42.
- Weisstein, E. (2006). Gamma function. From MathWorld—A Wolfram Web Resource.

- Wiener, N. (1949). *Extrapolation, Interpolation and Smoothing of Stationary Time Series*. Cambridge, MA: MIT Press.
- Wilks, S. (1938). The large sample distribution of the likelihood ratio for testing composite hypothesis. *Annals of Mathematical Statistics* 9, 60–62.
- Willems, F., Y. Shtarkov, and T. Tjalkens (1995). The context-tree weighting method: basic properties. *IEEE Transactions on Information Theory* 41, 653–664.
- Woodroffe, M. (1982). On model selection and the arcsine laws. *Annals of Statistics* 10, 1182–1194.
- Xie, Q., and A. Barron (1997). Minimax redundancy for the class of memoryless sources. *IEEE Transactions on Information Theory* 43, 646–657.
- Xie, Q., and A. Barron (2000). Asymptotic minimax regret for data compression, gambling and prediction. *IEEE Transactions on Information Theory* 46(2), 431–445.
- Yamanishi, K. (1998). A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Transactions on Information Theory* 44(4), 1424–1439.
- Yamazaki, K., and S. Watanabe (2003). Singularities in mixture models and upper bounds of stochastic complexity. *International Journal of Neural Networks* 16, 1029–1038.
- Yang, Y. (2000). Mixing strategies for density estimation. *Annals of Statistics* 28(1), 75–87.
- Yang, Y. (2005a). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* 92(4), 937–950.
- Yang, Y. (2005b). Consistency of cross-validation for comparing regression procedures. Submitted for publication.
- Yang, Y., and A. Barron (1998). An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory* 44, 117–133.
- Yang, Y., and A. Barron (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics* 27, 1564–1599.
- Yu, B. (1994). Lower bound on the expected redundancy for classes of continuous Markov sources. In S. Gupta and J. Berger (Eds.), *Statistical Decision Theory and Related Topics*, volume V, pp. 453–466.

- Yu, B. (1996). Lower bounds on expected redundancy for nonparametric classes. *IEEE Transactions on Information Theory* 42, 272–275.
- Yu, B., and T. Speed (1992). Data compression and histograms. *Probability Theory and Related Fields* 92, 195–229.
- Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with  $g$ -prior distributions. In P. Goel and A. Zellner (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pp. 223–243. Amsterdam: North-Holland.
- Zhang, J., and J. Myung (2005). A note on informative normalized maximum likelihood with data prior. Manuscript in preparation.
- Zhang, T. (2004a). Learning bounds for a generalized family of Bayesian posterior distributions. In S. Thrun, L. K. Saul, and B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems*, volume 16. Cambridge, MA: MIT Press.
- Zhang, T. (2004b). On the convergence of MDL density estimation. In Y. Singer and J. Shawe-Taylor (Eds.), *Proceedings of the Seventeenth Conference on Learning Theory (COLT' 04)*, Lecture Notes in Computer Science. New York: Springer-Verlag.





## List of Symbols

- $\mathbf{0}$ : null vector in  $\mathbb{R}^l$ , 43  
 $\mathbf{1}$ : indicator function, 52  
 $\mathcal{A}$ : data alphabet, 6  
 $A_\lambda$ :  $\lambda$ -affinity, 645  
 $a(\theta)$ : luckiness function, 309  
 $\hat{\alpha}$  where  $\alpha \in \{P, k, \gamma, \theta, (\theta, k)\}$ :  
     two-part code estimator,  
     132  
 $\hat{\alpha}$ , with  $\alpha \in \{\theta, \mu, \beta\}$ : ML  
     parameter, 58  
 $\alpha^*$  where  $\alpha \in \{P, k, \theta, \mu, \beta, \gamma\}$ :  
     “truth”, 143  
 $\text{arg}$ : argument of min/max, 44  
 $\mathcal{B}$ : Bernoulli/Markov model, 58  
 $\mathcal{B}_{\mathbb{Q}}$ : rational Markov models, 185  
 $B(\epsilon)$ :  $\epsilon$ -ball, 44  
 $B_{I(\theta)}$ : Mahalanobis ball, 121  
 $B_{\text{kl}}$ : KL ball, 220  
 $\mathbb{B}$ : Booleans, i.e.  $\{0, 1\}$ , 6  
 $\beta$ : canonical parameter, 601  
 $\hat{\beta}$ , see  $\hat{\alpha}$   
 $\mathcal{C}$ : code, coding system, 80  
 $\text{CCOMP}^{(n)}$ : constrained  
     complexity, 302  
 $\text{COMP}^{(n)}$ : model complexity, 180  
 $\text{CUP}(\cdot, \cdot)$ : CUP universal model,  
     373  
 $\text{cov}$ : covariance, 48  
 $D$ : KL divergence, 104  
 $D$ : data sequence, 6  
 $D$ : KL div. rel. to  $\ln$ , 111  
 $\Delta^{(m)}$ :  $m$ -dimensional unit  
     simplex, 42  
 $D_{P^*}$ : extended KL divergence,  
     625  
 $\bar{d}_\lambda$ : Rényi divergence, 478  
 $d_\lambda$ : unnormalized Rényi  
     divergence, 645  
 $d, d(x^n)$ : parameter precision, 137  
 $\det$ : determinant of matrix, 43  
 $E$ : expectation, 48  
 $E_\theta$ : expectation under  $P_\theta$ , 57  
 $\text{EACOMP}^{(n)}$ : exponentiated  
     asymptotic complexity,  
     216  
 $f$ : probability density function, 46  
 $\bar{f}_{\text{IND}}$  with  $\text{IND} \in \{\text{meta, plug-in, nml, Bayes, Jeffreys, two-part}\}$ , see  $\bar{P}_{\text{IND}}$   
 $\phi$ : sufficient statistic, 601  
 $\text{FSS}$ : fitted sum of squares, 344  
 $\Gamma$ : Gamma function, 45  
 $\gamma$ : model index, 61  
 $\gamma^*$ , see  $\alpha^*$   
 $\mathcal{H}$ : hypoth. class of functions, 63  
 $H$ : (Shannon) entropy, 103  
 $H(\Theta_0, \epsilon)$ : hypercube fill up  
     number, 222  
 $H$ : entropy rel. to  $\ln$ , 111  
 $h^*$ , see  $\alpha^*$   
 $\text{He}^2$ : squared Hellinger distance,  
     478  
 $\text{He}_\lambda^2$ : generalized squared  
     Hellinger divergence,  
     645  
 $I$ : Fisher information, 112

- $\inf$ : infimum, 44  
 $J$ : observed information, 112  
 $K$ : Kolmogorov complexity, 571  
 $k$ : number of parameters, 61  
 $k^*$ , see  $\alpha^*$   
 $\ddot{k}$ , see  $\ddot{\alpha}$   
 $\kappa$ : kernel function, 390  
 $\mathcal{L}$ : set of codelength functions, 99  
 $L$ : codelength function (of code or coding system), 83  
 $L_U$ : length function of uniform code, 88  
 $\bar{L}_{\text{IND}}$  with  $\text{IND} \in \{\text{meta, plug-in, nml, Bayes, Jeffreys, two-part}\}$ , see  $\bar{P}_{\text{IND}}$   
 $\bar{L}$ : length function of universal code, 174  
 $\bar{L}_{\text{CUP}}$ : CUP univ. code, 409  
 $\dot{L}_n$ : first part of two-part code, 272  
 $\bar{L}_{\text{cond-2-p}}$ : conditional two-part codelength, 285  
 $\text{LREG}$ : luckiness regret, 311  
 $\ln$ : natural logarithm, 45  
 $\log$ : base 2 logarithm, 45  
 $\mathcal{M}$ : model, 57  
 $\mathcal{M}^{(k)}$ :  $k$ -dimensional model, 60  
 $\langle \mathcal{M} \rangle$ : information closure, 376  
 $\mathcal{M}_{\text{loc}}$ : normal location family, 298  
 $\ddot{\mathcal{M}}_n$ : discretization of  $\mathcal{M}$ , 476  
 $\mathcal{M}_\Theta$ : model indexed by  $\Theta$ , 57  
 $\mathcal{M}^{\mathbf{X}}$ : linear model, 349  
 $M(\Theta_0, \epsilon)$ : packing number, 222  
 $\mu$ : mean-value parameter, 610  
 $\hat{\mu}$ , see  $\hat{\alpha}$   
 $\max$ : maximum, 44  
 $\min$ : minimum, 44  
 $N(\Theta_0, \epsilon)$ : covering number, 222  
 $\mathbb{N}$ : normal distribution, 51  
 $\mathbb{N}$ : natural numbers, 42  
 $O, o, \text{ORDER}$ : order notation, 45  
 $\mathcal{P}^{(k)}$ : polynomial linear regression model, 64  
 $P$ : probability distr./source/mass fn., 46  
 $P(\cdot | \cdot)$ : conditional prob., 50  
 $P^{(n)}$ : marginal on  $\mathcal{X}^n$ , 54  
 $\bar{P}_{\text{Bayes}}$ : Bayesian univ. model, 175  
 $P^*$ , see  $\alpha^*$   
 $\ddot{P}$ , see  $\ddot{\alpha}$   
 $\bar{P}_{\text{Cnml-1}}$ : conditional NML universal model, 322  
 $\bar{P}_{\text{Cnml-2}}$ : conditional NML universal model, 321  
 $\bar{P}_{|n}$ : prequential estimator, 463  
 $\bar{P}_{\text{Jeffreys}}$ : Bayesian universal model with Jeffreys' prior, 235  
 $\bar{P}_{\text{LB}}$ : Liang-Barron universal model, 325  
 $\bar{P}_{\text{Lnml-1}}$ : luckiness LNML universal model, 309  
 $\bar{P}_{\text{Lnml-2}}$ : luckiness LNML universal model, 311  
 $\bar{P}_{\text{lucky-Jeffreys}}$ : luckiness Bayesian universal model, 313  
 $P_{\text{me}}$ : maximum entropy distribution, 638  
 $\bar{P}_{\text{meta}}$ : metauniversal model, 304  
 $P_{\text{mre}}$ : minimum relative entropy distribution, 640  
 $\bar{P}_{\text{nml}}$ : NML universal model, 181  
 $\bar{P}_{\text{plug-in}}$ : plug-in univ. model, 197

- $\bar{P}_{\text{top}}$ : top-level univ. model, 406  
 $\mathbb{P}$ : empirical distribution/relative frequency, 53  
 $\pi$ : prior density, 493  
 $\psi$ : cumulant generating function, 601  
  
 $r$ : carrier density, 601  
 $\mathbb{R}$ : real numbers, 41  
RED: redundancy, 177  
 $\text{RED}_{\text{max}}$ : worst-case redundancy, 177  
REG: regret, 179  
 $\text{REG}_{\text{In}}$ : regret in “nats”, 240  
 $\text{REG}_{\text{max}}$ : worst-case regret, 180  
RISK: KL risk, 467  
 $\overline{\text{RISK}}$ : Césaro KL risk, 467  
RSS: residual sum of squares, 343  
  
 $S^{\mathbf{X}}$ : linear model, 349  
 $\Sigma$ : covariance matrix, 354  
SSE: sum of squared errors, 341  
sup: supremum, 44  
  
 $\top$ : transpose, 43  
 $\Theta$ : set of parameter values, 57  
 $\Theta_{\text{mean}}$ : mean-value parameter space, 610  
 $\Theta_{\text{can}}$ : canonical parameter space, 601  
 $\Theta_n$ : discretized parameter set, 274  
 $\Theta^{(k)}$ :  $k$ -dimensional parameter space, 60  
 $\Theta^{[+\delta]}$ :  $\Theta$  “blown up” by  $\delta$ , 248  
 $\ddot{\Theta}_d^{(k)}$ : discretized parameter set, 137  
 $\hat{\theta}_{\text{map}}$ : MAP estimator, 493  
 $\hat{\theta}_{\text{mean}}$ : Bayes mean estimator, 495  
 $\hat{\theta}_{\text{wf}}$ : Wallace-Freeman estimator, 497  
  
 $\hat{\theta}_\gamma$ : ML estimator in  $\mathcal{M}_\gamma$ , 372  
 $\hat{\theta}_a$ : LML estimator rel. to  $a(\theta)$ , 311  
 $\hat{\theta}$ , see  $\hat{\alpha}$   
 $\hat{\theta}$ , see  $\hat{\alpha}$   
 $\hat{\theta}^{(k)}$ : ML estimator in  $\mathcal{M}^{(k)}$ , 60  
 $\hat{\hat{\theta}}$ : discretized ML estimator, 224  
 $\theta$ : parameter vector, 57  
 $\hat{\theta}_{\alpha, \mu_0}$ : modified ML estimator, 260  
  
var: variance, 48  
  
 $W, w$ : prior distr./density, 74  
 $w$ : grid width, 137  
 $w_{\text{Jeffreys}}$ : Jeffreys’ prior, 234  
 $w_{\text{lucky-Jeffreys}}$ : luckiness Jeffreys prior, 313  
  
 $\mathcal{X}$ : sample space, 46  
 $\mathcal{X}^+, \mathcal{X}^*$ , 53  
 $\mathbf{X}$ : design matrix, 340  
 $\mathbf{x}$ : data vector, 319  
 $x^0$ : empty sample, 53  
  
 $\mathbf{y}$ : data vector, 340  
  
 $Z$ : noise term, 21  
 $Z(\beta)$ : partition function, 601  
 $\mathbb{Z}$ : integers, 42  
  
 $\lceil \cdot \rceil$ : ceiling, 45  
 $\sim$ : distributed as, 47  
 $\rightsquigarrow$ : ratio tends to 1, 45  
 $\uparrow$ : “undefined”, 47  
 $|\cdot|$ : number of elements in set  $\cdot$ , 45  
 $|x|$ : absolute value of  $x$ , 45  
 $\propto$ : proportional to, 45  
 $:=$ : “is defined as”, 42



## Subject Index

- a priori expected codelength, 555
- absolute singularity, 507
- accidental information, *see*  
information, accidental
- action, 574
- adaptive criterion, 417
- adaptivity, 193
- admissible, 527
- affinity, 499, **645**
- agnostic, 579
- AIC, 417, 532, 541, 549  
and BIC and MDL, 552  
regression, 450
- Akaike information criterion, *see*  
AIC
- algorithmic  
MDL, *see* MDL, idealized  
randomness, 11  
statistics, 11
- almost in-model, 493
- $\alpha$ -factor, 479, 512
- $\alpha$ -two-part code MDL, 477, 503,  
512, 645
- alphabet, **6**, 79, 80
- annealed entropy, 583
- apology, 561
- approximation, 376
- AR process, 214
- arbitrariness, 29, 152
- ARMA process, 214, 508
- astronomy, 545
- asymptotic, 429  
Bayesian code regret, 232, 244  
CNML regret, 323  
distinguishability, *see*  
distinguishability,  
asymptotic  
linear model regret, 366  
LNML regret, 312  
parametric complexity, 211  
plug-in code regret, 260  
two-part code regret, 273
- average, 631  
and approximation, 453  
Césaro, 474  
case vs. worst-case, 451
- averaging, 26, 341, 346
- ball, 44  
distinguishability, 221  
Kullback-Leibler, 220  
Mahalanobis, 121
- Barron, A.R., 242
- basis functions, 336
- batch, 419
- Bayes  
factors, 418, 539, 540, 549  
predictive interpretation,  
541  
formula, **50**, 75, 192  
generalized, 317  
nonparametric, 543  
vs. MDL, 533
- Bayesian  
brands, 544  
confirmation theory, 545  
estimator, *see* estimator,  
Bayesian  
evidence, 77, 175  
hypothesis testing, 650  
inconsistency, 543

- Information Criterion, *see*
  - BIC
  - interpretation, 418
  - linear model, *see* linear model, Bayesian
  - MAP, *see* MAP
  - marginal likelihood, 77
  - mixture, 77, 175
  - predictive distribution, *see* distribution, predictive
  - principles, 532
  - regret, 232
  - statistics, *see* statistics, Bayesian
  - universal model, *see* universal model, Bayesian
- Bernoulli family, *see* model, Bernoulli
- beta prior, 258
- BIC, 417, 532, 549
  - and AIC and MDL, 552
  - regression, 450
- binary tree, 92
- biological applications, 492
- Booleans, 42
- bound
  - Chernoff, 634, 636
  - generalization, 586
  - grand Cramér-Rao, 455
  - Hoeffding, 634, 636
  - Rissanen, 454
- boundary, 44
- bounded set, 44
- Brandeis dice, 638, 643
- Césaro
  - average, 474
  - consistency, *see* consistency, Césaro
  - KL risk, *see* KL risk, Césaro
  - universal model, 474
- calculation, *see* computation
- canonical
  - parameterization, *see* parameterization, canonical
  - prior, *see* prior distribution, canonical
- carving up models, 160
- central limit theorem, 56, 113, 220, 247
  - and relative entropy, 129
  - multivariate, 633
  - uniform, 289
- chain rule, *see* probability, chain rule
- cheating, 571, 572
- Chernoff
  - bound, 634, 636
  - divergence, 646
  - information, 650
- $\chi^2$ -divergence, 517
- choose function, 127
- classification, 72, 439
- Clinton, W.J., 437
- closure, 44
- CLT, *see* central limit theorem
- clustering, 229, 407
- CNML, *see* conditional NML
- code, 7, **80**
  - comma-free, 86
  - complete, 90, 94, 99
  - concatenation, 84, 86
  - conditional, 86, 290, 296, 302, 316, 433, 572
  - Liang and Barron, 446

- Liang-Barron, 325, 449, 470
  - design, 152, 157, 158
  - efficient, 90, 94
  - fixed-length, 88
  - for Markov chains, 137
  - inefficiency, 161
  - instantaneous, **84**
  - meta-two-part, *see*
    - meta-two-part code
  - metauniversal, *see*
    - metauniversal code
  - minimax, 99
  - NML, *see* NML
  - one-part, 152
  - optimal, 135
  - partial, 80
  - predictive, 152
  - predictive MDL, 198
  - prefix, 83, **84**
  - quasi-uniform, 90, 100, 159, 187, 422, 425
  - Shannon-Fano, **95**, 110, 136
  - Shtarkov, *see* NML
  - standard for integers, 101, 137, 186
  - trivial, 89
  - two-part, *see* two-part code
    - design, 161
  - uniform, 87, 99, 106, 137, 285
  - uniquely decodeable, 7
  - universal, *see* universal code
  - word, 80, 82
- codelength
  - absolute, 305, 568
  - as probability, 96
  - difference, 411
  - excess, 177
  - expected, 625, 630, 642
  - function, **99**
  - ignoring, 439
  - interpretation, 433
  - invariance, 98
  - minimax, 87, 106, 624, 637, 641
  - minimax relative, 643
  - noninteger, 95
  - observed average, 630
  - relative, 305, 568
  - Shannon-Fano, 114
  - worst-case, 87
- coding
  - by index, 284
  - model index, 424
  - system, 7, **80**
    - lossy, 80
    - partial, 80
    - singular, 80
  - with help of, 136
- coherent, 544
- compactness, 44
- compatibility condition, 53, 194
- compatible
  - prior distribution, 232
- complete code, *see* code, complete
- complexity, 135, 143, 180, 411, 416
  - adjusted, 427
  - and volume, 216, 222
  - Bernoulli model, 212, 227
  - classification, 582
  - computation, *see*
    - computation, parametric
    - complexity
  - conditional, 302
  - constrained, 302, 440
  - data-dependent, 584
  - distinguishability, 224
  - exponentiated asymptotic, 216

- for exponential families, 226
- histogram, 378
- histogram model, 228
- infinite, 215, 295, 420
- Kolmogorov, 8, 546, 570
- learning theory, 582
- multinomial model, 228
- nonasymptotic, 228
- of hypotheses, 30
- of models, 30
- PAC-Bayesian, 589
- parametric, 208, 210
  - asymptotic, 211
- Rademacher, 583
- renormalized, *see* RNML
- second order, 307
- simplification, 226
- stochastic, *see* stochastic
  - complexity
- composite hypotheses, 421
- composition of experiments, 345
- compression, 73, 413, 595
  - and fast learning, 469, 482
  - and regularity, 103
  - game, 643
  - interpretation, 415
  - lossy, 571
  - schemes, 584
- computable distribution, 546, 571
- computation
  - and Bayes, 538
  - parametric complexity, 226, 422
  - plug-in codelength, 269
  - two-part code, 150
  - universal codelengths, 428
- computational learning theory, *see*
  - learning theory,
  - computational
- concave function, 42
- concentration, 512
- concept learning, 72
- conditional
  - code, *see* code, conditional
  - description method, 86
  - distribution, 49
  - exponential family, 619
  - limit theorem, 127, 569
  - NML, 320, 470
    - and Jeffreys, 323
    - and luckiness NML, 322
    - asymptotic regret, 323
    - linear model, *see* linear regression, NML
    - model selection, 431, 446
  - NML-1, 322, 420, 426
  - NML-2, 321, 420, 426
  - NML-3, 323
  - probabilistic source, *see*
    - probabilistic source,
    - conditional
  - regret, 365
  - two-part, *see* two-part code, conditional
  - universal code, *see*
    - code, conditional
  - universal model, *see* code, conditional
- confidence, 411, 421, 535
  - level, 413, 585
- consistency, 71, 73, 143, 145, 153, 425, 449, 477
  - AIC, 550
  - and measure 0, 506
  - BIC, 550
  - Césaro, 467, 472
  - discussion, 501
  - essential, 468, 472



- information, 467
- KL, 467, 472
- linear regression, 508
- MDL
  - general, 514
  - MDL model selection, *see* MDL model selection, consistency
  - misspecification, 504
  - model selection, 74
  - of ML estimator, 59
  - peculiarities, 511, 593
  - prequential MDL, *see* prequential MDL, consistency
  - scenarios, 501
  - theorem, 155, 467, 478, 506
  - trivial, 504
  - two-part MDL, *see* two-part MDL, consistency
  - weak, 479, 503
- constrained parametric
  - complexity, *see* complexity, constrained
  - complexity, constrained
- constraints, 568
- continuum limit, 488, 497
- convergence rate, 59, 158, 161, 387, 425, 467, 477, 501, 506, **515**
- AIC, 550
- BIC, 550
- essential, 473
- MDL, 520
  - model selection, 522
- minimax, **519**
- parametric, 469, 483
- prequential MDL, 467
- scenarios, 501
- two-part MDL, 478
  - uniform, 516
- convex, 42
  - analysis, 604
  - duality, 613
  - function, 42
  - set, 42
- correlated prediction error, 567
- countable, 423
  - universal model, 184
- counting interpretation, 416
- covariance, 48, 606, 634
  - linear model, 353
- covering number, 222, 281
- Cramér's Theorem, 634
- Cramér-Rao bound, 455
- critical value, 412
- cross-product matrix, *see* Fisher information, regression matrix
- cross-validation, 541, **565**
- crude MDL, *see* MDL, crude
- Csiszár-Shields, *see* inconsistency, Csiszár-Shields
- cumulant, 605
- cumulative risk, 467
- CUP, 65
  - codes, *see* universal model, CUP
  - model class, *see* model class, CUP
- curvelength, 218
- CV, *see* cross-validation
- data, **6**, 69
  - alphabet, *see* alphabet item, 6
  - nondegenerate, 142
  - sample, 6
  - sequence, 6

- unseen, 419, 567
- virtual, 347
- Dawid, A.P., 190, 261, 562
- De Finetti, B., 548
- decision
  - function, 69
  - theory, 594
- decoder, 83, 172
- decoding function, 83
- defective, 94, 103, 149
- degree-of-belief, 539
- denoising, 349, 407, 423
- density
  - differentiable, *see*
    - differentiable density
  - estimation, 407, 460, 464
    - histogram, *see* histogram
    - density estimation
  - nonparametric, *see*
    - nonparametric density
  - estimation
  - function, 46
- description method, 7, 80
  - conditional, 86
  - prefix, 83, 84
  - quasi-uniform, 90, 100
- descriptions, 83
- design
  - matrix, 340, 391
  - principle, 29, 525, 591
- determinant, 43
- deviance, 539
  - information criterion, 539
- DIC, 539
- dice, 638, 643
- diffeomorphism, 611
- differentiable density, 370, 371, 503, 520, 525
- differential entropy, 104
- dimensionality, 219
- Dirichlet distribution, 263
- discrete estimator, 489
- discretization, 104, 137, 278, 485
- distance
  - Euclidean, *see* Euclidean
  - distance
  - Hellinger, *see* Hellinger
  - distance
  - integrated squared error, 517
  - Mahalanobis, *see*
    - Mahalanobis distance
  - relations, 517
- distinguishability, 153, 182, 216, 339, 416
  - and complexity, 224
  - and conditional two-part code, 290
  - and Jeffreys' prior, 236
  - and KL divergence, 219
  - and phase transition, 290
  - asymptotic, 153, 507
  - ball, 221
  - level, 224, 279
  - number of distributions, 224
  - region, 220
- distinguishable distributions, *see* distinguishability
- distribution
  - Bernoulli, *see* model, Bernoulli
  - beta, 258
  - computable, 546, 571
  - conditional, 49
  - Dirichlet, 263
  - distinguishable, *see* distinguishability
  - gamma, 362, 445
  - generating, 71

- joint, 49
- marginal, 49, 55
  - Bayesian, 77
- multivariate normal, 50, 634
- normal, *see* model, normal
- posterior, *see* posterior distribution
- predictive, 77, 391, 460, 461
  - linear model, 356
- prior, *see* prior distribution
- product, 51, 55
- square-root inverted gamma, 362, 445
- true, 20, 525
- uniform, 99, 638
- divergence
  - $\chi^2$ , 517
  - KL, *see* KL divergence
  - Rényi, *see* Rényi divergence
  - relations, 517
- DNA, 492
- d*-risk, 515
- Dutch book, 544
- early stopping, 490
- efficient code, *see* code, efficient
- eigenvalues, 43
- eigenvector, 120, 278
- ellipsoid, 120, 278
- Ellsberg paradox, 545
- EM algorithm, *see* expectation-maximization
- empirical
  - error, 581
  - loss, 582
  - Rademacher complexity, 583
  - risk minimization, 580
- encoder, 83, 172
- end-of-input marker, 86
- ensemble, 631
- entropification, 574, 588
- entropy, 103, 615
  - annealed, 583
  - coding interpretation, 104
  - combinatorial interpretation, 127
  - differential, 104
  - for exponential family, 606
  - maximum, 106, 600, 624, 637
    - principle, 567
    - prior, 569
  - minimum relative, 624
  - of normal distribution, 640
  - Rényi, 583
  - Rényi vs. Shannon, 650
  - relative, *see* KL divergence
  - Shannon, 104
- epigraph, 42
- equalizer strategy, 181
- ERM, 580
- essential
  - consistency, *see* consistency, essential
  - convergence, *see* convergence rate, essential
- estimation
  - density, *see* density estimation
  - maximum likelihood, *see* likelihood, maximum
  - nonparametric, *see* nonparametric density estimation
  - parameter, 70, 476
    - MDL, *see* MDL parameter estimation
  - predictive MDL, *see* prequential MDL

- estimator, 57
  - $\alpha$ -two-part MDL, 477, 503, 512, 645
  - Bayes MAP, *see* MAP
  - Bayes mean, 354, 495, 556
  - Bayesian, 268
  - discrete, 489
  - in-model, 263, 268, 462, 463, 491, 576
  - Krichevsky-Trofimov, 258
  - Laplace, 258, 527
  - least-squares, *see* least-squares
  - maximum likelihood, *see* likelihood, maximum
  - model selection-based, 509
  - out-model, 263, 268, 463, 485, 491
  - SMML, *see* minimum message length, strict superefficient, 455, 527
  - two-part MDL, *see* two-part MDL
  - unbiased, 268
  - Wallace-Freeman, 497, 560
- Euclidean distance, 43, 159, 515, 517
  - rescaled, 120
- evolutionary trees, 492
- exchangeable, 434
- expectation, 48, 52
  - vs. hope, 533, 535
- expectation-based MDL, *see* MDL, expectation-based
- expectation-maximization, 151, 427, 490
- expected
  - codelength, 625
  - redundancy, *see* redundancy, expected
- experiment composition, 345
- experimental design, 32
- experts, 574
- exponential family, 66, 67, 599, 600
  - and maximum entropy, 638
  - canonical parameterization, 287
  - conditional, 619
  - discrete, 286
  - entropy, 606
  - Fisher information, 606
  - i.i.d., 603, 619
  - linear model, 352
  - mean, 606
  - minimal, 66, 601
  - ML estimator, 629, 630, 632
  - of general sources, 617
  - robustness property, 208, 240, 266, 605, **624**, 641
  - variance, 606
- exponentiated asymptotic complexity, 216
- extended KL divergence, 625
- family
  - exponential, *see* exponential family
  - likelihood ratio, 647
  - normal, *see* model, normal
- feature vector, 336
- Fechner's model, *see* model, Fechner's
- Fisher information, 607
- Fisher information, 120, 121, 221, 275, 606, **619**, 634
  - cancellation, 488

- determinant, 211
- empirical, 112
- expected, 119, 241
- normal models, 300
- Observed, 620
- observed, 112, 241
- regression matrix, 343, 350
- Fisher, Sir Ronald, 117, 145
- fitted sum of squares, 344
- footing, 416, 425
- forward validation, 448, 563
- Foster and Stine approach, 329
- France, 75
- F*-ratio, 446
- free parameters, 212
- frequency, 59, 61, 285
  - in sample, 53
- frequentist statistics, *see* statistics, frequentist
- function
  - concave, 42
  - convex, 42
  - decoding, 83
  - encoding, 80
  - partial, 47
- functional form, 24, 216
- gambling
  - Kelly, 98, 191
- game
  - compression, 643
  - theory, 573
  - zero-sum, 637
- gamma
  - distribution, 362, 445
  - function, 45
- gappy sequence, 473
- Gauss curve, 113
  - multivariate, 117
- Gauss, C.F., 145
- Gaussian, *see* model, normal
  - prior, *see* prior distribution, Gaussian
  - process, 364, 371, 390, 394, 472, 504, 542
    - as universal model, *see* universal model, Gaussian process
  - finite-dimensional, 399
  - regression, 395
- generalization
  - bound, 586
  - error, 579
  - performance, 73, 581
- generalized
  - Bayes, 317
  - linear model, *see* linear model, generalized
  - NML, *see* luckiness NML-1
- geometric family, *see* model, geometric
- Germany, 75
- Gibbs' inequality, 105
- global
  - maximum, 151
  - ML principle, 149, 532
- gMDL, 443, 449
- goals of inductive inference, 71
- goodness-of-fit, 135, 143
- g*-prior, 444
- gradient, 116
  - descent, 427
- gzip, 537, 538
- heavy tails, 451
- Hellinger
  - affinity, 499, 645, 646

- distance, 71, 478, 491, 510, 512, 515, 517, 645
- risk, 490, 504
- Hessian matrix, *see* matrix, Hessian
- Hilbert space, 394, 396
- hill-climbing, 151
- hindsight, 174, 419
- histogram, 482
  - Bayes, 378
  - density estimation, 376, 471, 510
    - alternative, 471
    - irregular, 471
    - minimax optimal, 382
    - model, 376
    - NML, 378
    - regular, 377
    - universal model, 378
- Hoeffding bound, 634, 636
- honest, 161, 567
- hope, 533, 535
- horizon, 190, 194, 563
- hypercube, 42, 222
- hyperparameter, 441
- hypothesis, **13**
  - complexity, 30
  - composite, **69**, 72
  - individual, 69
  - point, **13**, **69**
  - selection, 70, 406
  - simple, 70
  - singleton, 69
  - testing, 412
    - and luckiness, 421
- I-divergence, 646
- i.i.d., 54, 62
- idealized MDL, *see* MDL, idealized
- ignoring codelength, 439
- iMDL, 448, 449
- improper prior, 317
- in-model estimator, *see* estimator, in-model
- incompleteness, 274
- incompressibility, 103
- inconsistency, 146, 504, 530, 590, 593
  - Bayesian, 543
  - Csiszár-Shields, 506, 513
- independence, 51, 54, 638
- index coding, 424
- index of resolvability, 483
- indicator function, 52
- individual-sequence, 178, 564
  - MDL, *see* MDL, individual sequence
  - prediction, 573
  - prequential MDL estimator, 464
- inductive inference, 69
  - goal of, 71
- ineccsi, 301
  - model, 209
  - sequence, 210
  - subset, 209
- inefficiency of codes, 161
- inequality
  - Gibbs', 105
  - information, 101, 103, 105, 118, 452
  - Jensen's, 105, 475, 500
  - Kraft, 91, 500
  - Markov's, 56
  - no-hypercompression, **102**, 155, 413, 535

- triangle, 480
- infimum, 44
- infinite
  - complexity, *see* complexity,
  - infinite
  - horizon, 563
- infinite-dimensional, 394
- infinitely many models, 422
- infinity problem, 295
- information
  - accidental, 416
  - Chernoff, 650
  - closure, 376, 471
  - consistency, 467
  - criterion
    - AIC, *see* AIC
    - BIC, *see* BIC
  - inequality, *see* inequality,
  - information
  - matrix, *see* Fisher information
  - meaningful, 416
  - observed, 112
  - source, 53
  - theory, 72, 79
- inner product, 43
- input vector, 336
- insight, 492
- integers, 42
- integral
  - Gaussian, 240
- integrated squared error, 517
- integration
  - Laplace method, 239
  - saddle point method, 239
- interior, 44
- interval
  - closed, 42
  - open, 42
  - unit, 42
- invariance theorem, 10
- irregular histogram, 471
- Jacobi matrix, 611
- Jeffreys' prior, *see* prior
  - distribution, Jeffreys'
- Jensen's inequality, *see* inequality,
  - Jensen's
- joint distribution, 49
- Kelly gambling, *see* gambling,
  - Kelly
- kernel, 472, 542
  - density estimator, 371
  - finite-dimensional, 398
  - function, 390
  - Matern, 400
  - Mercer, 400
  - polynomial, 391
  - RBF, 393
  - trick, 393
- kernelization, 390
- KL consistency, 467
- KL divergence, 103, 104, 120, 154,
  - 156, 159, 201, 481, 515,
  - 517, 620, 625, 633, 634,
  - 647
  - and distinguishability, 219
  - ball, 220
  - chain rule, 465
  - coding interpretation, 105
  - extended, 266
  - linear model, 351
  - robustness, *see* exponential
    - family, robustness
    - property
  - Taylor expansion, 117, 240,
  - 266, 276, 518, 636
- KL risk, 466

- Césaro, 467, 474, 479, 520
- Kolmogorov
  - complexity, *see* complexity, Kolmogorov
  - minimum sufficient statistic, 11, 571
  - structure function, 571
- Kolmogorov, A.N., 8
- Kraft inequality, *see* inequality, Kraft
- Krichevsky-Trofimov estimator, 258
- Kullback-Leibler divergence, *see* KL divergence
- label, 579
- Laplace
  - integration, 239
  - rule of succession, 258
- Laplace, P.S. de, 258
- large deviations, 624, 634
- law
  - of iterated logarithm, 247
  - of large numbers, 55, 125
- learning and compression, 469, 482, 595
- learning theory
  - computational, 72, 573, 579
  - luckiness in, 309
  - statistical, 72, 449, 525, 579
- least-squares, 337, **340**
  - and projection, 342
  - penalized, **346**, 355
  - predictive, *see* PLS
- leave-one-out
  - cross-validation, *see* cross-validation
  - error, 566
- Legendre transform, 614, 648
- length function, *see* codelength, function
- level of distinguishability, 224
- Liang-Barron, *see* code, conditional, Liang-Barron
- light tails, 480
- likelihood, 57, 111
  - expected, 117
  - maximum, 57, 111, 124, 527, 624, 629, 632
  - Bernoulli, 258
  - consistency, 59
  - discretized, 224
  - global, 149
  - linear model, 350
  - local, 427
  - luckiness, 311, 419, 484, 487, 494, 497, 498
  - modified, 258, **260**
  - principle, 144
  - vs. MDL, 147
- maximum normalized, *see* NML
- maximum renormalized, *see* RNML
- ratio, 631
  - family, 647
  - test, 412, 417
- linear
  - model, 64, 335, 337, **348**, 423, 428, 589, 617
  - and normal, 338, 349
  - as exponential family, 352
  - Bayesian, 354, 364, 390
  - covariance, 353
  - generalized, 401, 443, 449, 508
  - Jeffreys prior, 359



- KL divergence, 351
- MAP, 354
- marginal density, 355
- ML estimator, 350
- parameterization, 352
- posterior, 355
- predictive distribution, 356
- regret, 350
- regression, 63, **335**, 580, 617, 619
  - CNML, 446
  - consistency, 503, 508
  - gMDL, 443
  - iMDL, 448
  - Liang and Barron, 446
  - model selection, **438**
  - NML, 363
  - plug-in code, 448
  - RNML, 439
  - universal model, 363
- link function, 401
- LLN, *see* law of large numbers
- LML, *see* likelihood, maximum
  - luckiness
- LNML, *see* luckiness NML
- local maximum, 151
  - and NML, 427
- log
  - likelihood, 57
  - loss, *see* loss, logarithmic
  - score, *see* loss, logarithmic
- log-convex mixtures, 645
- log-likelihood
  - expected, 117
  - surface, 630
- logarithm, 45
- logistic
  - function, 401
  - regression, 630
- LOO, *see* leave-one-out
- loss, 574
  - 0/1, 540, 574
  - classification, 574
  - empirical, 582
  - function, 73
  - general, 461, 464
  - logarithmic, 190, 460, 574
  - simple, 575
  - squared, 337, 574, 579
  - symmetric, 575
- lossy compression, 571
- luckiness, 424, 433, 585
  - and Bayes, 310, 534
  - and hypothesis testing, 421
  - and PAC-Bayes, 586
  - Bayesian universal model, 313
  - coding, 296
  - function, 309, 354, 398, 406, 443, 483
    - choosing, 534
  - Gaussian process, 397
  - in learning theory, 309
- ML, *see* likelihood, maximum
  - luckiness
- NML, 309, 422
  - and conditional NML, 322
  - asymptotic regret, 312
  - linear model, *see* linear regression, NML
- NML-1, 309, 426
  - and RNML, 443
- NML-2, 311, 426, 485
- principle, 92, 159, 305, 420, 425, 449, 532
- rationale, 535
- regret, 311, 397, 422, 484, 493
- tilted Jeffreys' prior, 313, 484

- uniform, 485
- universal model, 308
  - Bayesian, 313
- machine learning, 69, 72
- Mahalanobis distance, 120, 278, 344, 350, 398, 517
- many-to-one, 42
- MAP, 311, 493, 556
  - and LML, 494
  - linear model, 354
- marginal
  - distribution, *see* distribution, marginal
  - likelihood, 175
- Markov chain, *see* model, Markov
- Markov's inequality, 56
- Matern kernel, 400
- matrix
  - cross-product, *see* Fisher information, regression matrix
  - design, 340
  - determinant, 43
  - eigenvalues, 43
  - Hessian, 43, 117
  - information, *see* Fisher information
  - inverse, 43
  - positive definite, 43, 120, 606
  - projection, 342
- MaxEnt, *see* entropy, maximum
- maximum, 44
  - a posteriori, *see* MAP
  - entropy, *see* entropy, maximum
  - likelihood, *see* likelihood, maximum
- probability principle, 95, 149, 532
- MDL
  - algorithmic, *see* MDL, idealized
  - and cheating, 571
  - and MML, 558
  - and prediction, 578
  - and traditional statistics, 469, 482
  - application, 38, 68, 592
  - consistency, *see* consistency
  - convergence rate, *see* convergence rate, MDL
  - criterion, 553
  - crude, 133, 389
  - decision theory, 594
  - expectation-based, 21, 407, 504, 530
  - idealized, 11, 30, 546, 570
  - individual-sequence, 21, 407, 484, 504, 523, 530
  - justification, 152
  - meta-principle, 160
  - model selection, 409
    - and null hypothesis testing, 413
  - Bayesian interpretation, 418
  - compression
    - interpretation, 415
  - consistency, 415, 505, 510
  - consistency theorem, 506
  - convergence rate, 522
  - counting interpretation, 416
  - definition, 427
  - discussion, 448
  - four interpretations, 415

- general, 410, **420**
    - gMDL, 443
    - iMDL, 448
    - infinitely many models, 422
    - Liang and Barron, 446
    - linear regression, **438**
    - nonparametric, 508
    - plug-in codes, 431
    - prequential interpretation, 419
    - refined, **426**
    - RNML, 439
    - simple, 411
  - model selection-based
    - estimator, 509
  - nonparametric regression, 448
  - nonprobabilistic, 20
  - parameter estimation, 483
    - approximation, 486
    - vs. LML, 487
  - philosophy, 199, 436, 487, 595
  - predictive, *see* prequential
    - MDL
  - prequential, *see* prequential
    - MDL
  - principle, 595
  - probabilistic, 20
  - problems, 593
  - refined, 17, 406
  - two-part code, *see* two-part
    - MDL
  - vs. Bayes, 533
  - vs. ML, 147
- mean, 48
- mean-value parameterization, *see*
  - parameterization,
  - mean-value
- meaningful information, *see*
  - information, meaningful
- medical testing, 75
- Mercer kernel, 400
- Mercury, 545
- messages, 83, 172
- meta-Bayes code, 374, 379, 389, 406, 471
- meta-MDL principle, 160
- meta-two-part code, 303, 373, 379, 389, 406, 409, 443, 471, 505
- metauniversal code, 296, **301**
  - problems, 308
- meteorology, 572
- minimal representation, 601
- minimax
  - analysis, 451
  - codelength, *see* codelength, minimax
  - convergence rate, *see* convergence rate, minimax
  - nonparametrics, *see* redundancy, nonparametric
  - optimal histogram code, 382
  - regret, *see* regret, minimax theorem, 644
- minimum, 44
  - message length, 493, 555
    - and MDL, 558
    - strict, 497, 556
  - relative entropy, *see* entropy, minimum relative, 640
- misspecification, 265, 502, 504, 530, 590, 593
- mistake probability, 182
- mixability, 576, 590

- mixture, 576
  - Bayesian, 77
  - family, *see* model, mixture
  - model, *see* model, mixture
- ML, *see* likelihood, maximum
- MML, *see* minimum message length
- model, **13**, 70
  - Bernoulli, **58**, 65, 66, 71, 118, 122, 233, 461, 491, 495, 602, 604, 607, 615, 629, 637
  - and Jeffreys, 236
  - complexity, 212
  - class, **13**, 70
    - CUP, 65, 370, **372**, 409, 502, 505, 556
    - fully nonparametric, 370
    - Gaussian process, *see* Gaussian process
    - NCUP, 372
    - nonparametric, 369, 468, 471, 492, 503
    - probabilistic, **60**, 70
  - complexity, 30, 180, 412
    - Bernoulli, 188, 227
    - multinomial, 228
    - nondecreasing, 189
    - Poisson, 189
  - conditional, 62
  - cost, 412
  - crazy Bernoulli, 217
  - exponential family, *see* exponential family
  - Fechner's, 23, 417, 505
  - functional form, 216
  - Gaussian process, *see* Gaussian process
  - geometric, 428, 603, 605, 607, 628
    - complexity, 299
  - histogram, *see* histogram
  - i.i.d., **62**, 433
  - linear, *see* linear model, 566
  - Markov, **60**, 65, 66, 71, 133, 185, 423, 513, 617, 618
    - code design, 158
    - hidden, 427, 437
  - meta-selection, 555
  - mixture, **68**, 427
  - multinomial, 59, 262, 378, 428, 602
    - complexity, 228
    - histogram, 228
  - naive Bayes, 562
  - nested, 24, **60**
  - non-nested, 23, 224, 431
  - normal, 65, 67, 264, 298, 428, 461, 602, 605, 626, 631, 640
    - and linear, 338, 349
    - complexity, 298
    - Fisher information, 300
    - mixture, **68**
    - RNML, 306
  - parametric, 57, 369, 375, 483, 519
    - Poisson, 65, 189, 428, 433, 534, 602, 604
      - complexity, 298
    - probabilistic, 57, 70
  - rich, 180
  - selection, 70, 406, **409**
    - by cross-validation, *see* cross-validation
  - CNML, 446
  - consistency, 74

- criterion, 509
  - MDL, *see* MDL, model selection
  - meta, 555
  - nested, 141
  - non-nested, 23, 141
  - prequential, *see* prequential model selection
  - warnings, 435
- Stevens's, 23, 417, 505
- time series, *see* time series
- trivial, 414
- true, 29
- universal, *see* universal model
- model selection-based estimator, 509
- multinomial, *see* model, multinomial
- multivariate normal distribution, *see* distribution, multivariate, normal
- mutual singularity, 154
- naive Bayes, 562
- nats, 110
- natural numbers, 42
- Nature, 644
- NCUP model class, 372
- Neal, R., 542
- nearest neighbor, 371
- nested, *see* model, nested
- neural networks, 72, 583
- Neyman-Scott problem, 527
- NMAP, 311
- nMDL, 439
- NML, 181, 183, 411, 412, 422, 513, 531
- conditional, *see* conditional NML
- generalized, *see* luckiness NML-1
- linear model, *see* linear regression, NML
- luckiness, *see* luckiness NML
- undefined, 183, 296, 298
  - geometric model, 299
  - normal model, 298
  - Poisson, 298
- no-hypercompression inequality, *see* inequality, no hypercompression
- noise, 416, 449, 575
- non-nested, *see* model, non-nested
- nondegenerate data, 142
- nondifferentiable density, 526
- nonineccsi sequences, 238
- noninformative prior, *see* prior distribution, noninformative
- nonmixability, 590
- nonparametric, 65, 369, 525
  - Bayes, 543
  - density estimation, 470
  - model class, *see* model class, nonparametric
  - model selection, 508
  - rate, 520
  - redundancy, *see* redundancy, nonparametric
  - regression, *see* regression, nonparametric
  - statistics, 35
- nonpredictive description, 492
- nonprequential, 196, 388
- nonuniform

- convergence rate, 516
  - universal model, 185
- normal distribution, *see* model, normal
- normal equations, 342
- normality rules, 357
- normalized
  - MAP, 311
  - maximum likelihood, *see* NML
- null
  - hypothesis, 412
  - model, 444, 514, 573
- objective
  - Bayes, 546
  - priors, 547
- observation, 6, 72
- Occam factor, 539
- Occam's razor, 29, 408, 539
  - as a methodology, 35
  - bound, 585
  - hidden, 542
- on-line decoding, 196
- one-to-one, 42
- online decoding, 135
- open problem, 264, 267, 268, 289, 299, 314, 323, 383, 413, 421, 451, 472, 490, 492, 511, 514, 522, 553, 569, 583
- order
  - dependence, 434
  - notation, 45, 83
- order notation, 83
- orthodox statistics, *see* statistics, frequentist
- out-model estimator, *see* estimator, out-model
- outlier, 451
- output vector, 336
- overfitting, 24, 133, 134, 145, 346, 416, 468, 490
- PAC-Bayes, 585
  - and luckiness, 586
  - and square root, 589
  - and SRM, 587
- PAC-MDL, 585
- packing number, 222, 281
- paradigm
  - Bayesian, 531
  - frequentist, 524
- parameter, 57, 284
  - estimation, *see* estimation, parameter, 141
  - precision, 283
  - space, 57
- parameterization, 57
  - canonical, 67, 211, 604, 611
  - dependence, 159, 307, 556
  - independence, 159, 211, 215, 485, 494
  - mean-value, 604, 611
  - properties, 615
  - uniform, 159
- parametric, 65
  - complexity, *see* complexity model, *see* model, parametric rate, 519
- partial
  - code, 80
  - function, 47
  - random variable, 47
- partition function, 601
- pattern recognition, 69, 72
- penalized least-squares, *see* least-squares, penalized

- permutation invariance, 328, 339, 435
- phase transition, 290
- Pitman-Koopman-Darmois, 604
- PLS, 448–450
- plug-in distribution, *see* universal model, plug-in
- point hypothesis, 406, 476
- point null hypothesis, 412
  - and luckiness, 421
- Poisson family, *see* model, Poisson
- polynomial, 70, 337, 341, 348, 391, 416, 437, 438
- positive (semi-) definite, *see* matrix, positive definite
- positive definite kernel, 393
- posterior
  - concentration, 512
  - convergence, 511
  - predictive, 78
- posterior distribution, 73, 75, 532
  - linear model, 355
  - meaning, 537
- pragmatic prior, *see* prior
  - distribution, pragmatic
- precision, 283
- prediction, 70, 191, 406, 459, 574
  - error, 73
  - individual-sequence, 573
  - probabilistic, 191
  - strategy, 190, 462
  - weather, *see* weather
    - forecasting
  - with expert advice, 574
- predictive
  - distribution, *see* distribution, predictive
  - least-squares, *see* PLS
  - MDL, *see* prequential MDL
- prefix
  - codes, 83
  - description methods, 83
- prequential, 364, 368, 388
  - analysis, 562
  - interpretation, 419
  - model selection, 419
  - parametric MDL, 484
  - plug in model, *see* universal model, plug-in
  - principle, 528
    - infinite horizon, 563
  - universal model, *see* universal model, prequential
- prequential MDL, 198, 460
  - consistency, 465, 502
  - consistency theorem, 467
  - estimation, 483
  - estimator, 461, 472
  - individual-sequence
    - estimator, 464
- prequentialization, 196
- principal axis theorem, 347
- prior, *see* prior distribution, 175
- prior density, *see* prior distribution
- prior distribution, 73, 74, 231, 532, 569, 585
  - $g$ -prior, 444
  - and luckiness, 534
  - beta, 258
  - canonical, 289, 292
  - compatible, 232
  - Diaconis-Freedman, 543
  - flat, 539
  - Gaussian, 354, 539
  - improper, 317, 420
  - informative, 420

- Jeffreys, 234, 258, 289, 418, 420, 447, 461, 513, 547, 570
  - and Bernoulli, 236
  - and boundary, 237, 244
  - and CNML, 323, 368
  - and distinguishability, 236
  - geometric, 300
  - linear model, 359
  - normal family, 299
  - Poisson, 300
  - Takeuchi-Barron
    - modification, 242
  - tilted, 312
  - undefined, 239, **299**
  - Xie-Barron modification, 239
- least informative, 235
- noninformative, 359
- objective, 547
- pragmatic, 534
- reasonable, 233
- reference, 547
- subjective, 556
- uniform, 258, 285, 379
- universal, integers, 101, 186, 423
- weakly informative, 539
- probabilistic source, 53, 69, 464, 575
  - and exponential families, 617
  - conditional, 62, 391, 588, 617, 619
- probability
  - 1-statement, 55
  - as codelength, 96
  - chain rule, 54, 465
  - conditional, 54
  - defective, **94**
  - density function, 46
  - mass function, 46
  - posterior, *see* posterior distribution
  - prior, *see* prior distribution
  - sequential decomposition, 54, 60, 465
- probably approximately correct, 586
- product distribution, 51
- projection, 342
- quadratic form, 344
- quasi-uniform, *see* code, quasi-uniform
- Rényi divergence, **478**, 491, 512, 645
  - interpretation, 649
  - unnormalized, 645
- Rademacher complexity, 583
- radial basis function, 393
- random
  - process, 53
  - sequence, 103
  - variable, 47
    - dependent, 51
    - independent, 51
  - vector, 47
- randomization, 432
- rate, *see* convergence rate
- RBF kernel, 393
- real numbers, 41
- receiver, 83, 172
- recipe, 504, 524
- redundancy, 177
  - almost sure, 199, 246, 265
  - conditional, 325, 446
  - CUP codes, 384



- expected, 199, 245, 325, 455, 467, 471
    - minimax, 247
  - histogram codes, 380
  - individual-sequence, 199
  - minimax, 202
    - expected, 247
  - nonparametric, 383, 387
  - relative, 265
  - stochastic, 199
  - terminology, 201
  - worst-case, 177
- reference code, 643
- refined MDL, 17, 406
- region of distinguishability, 220
- regression, 25, 63, 72, 336
  - Gaussian process, *see* Gaussian process
  - Gaussian process
    - linear, *see* linear regression
  - nonparametric, 448, 472
  - polynomial, *see* polynomial
  - ridge, 347
  - robust, 451
- regressor, 63, 336
- regret, **179**, 210
  - almost sure, 199, 244, 265
  - and maximum entropy, 568
  - asymptotic
    - Bayesian, 232
    - CNML, 323
    - linear model, 366
    - LNML, 312
    - NML, 211
    - plug-in, 260
    - two-part, 273
  - Bayes and linear model, 364
  - beyond log loss, 577
  - CNML and linear model, 365
  - conditional, 365
    - expected, 199, 202, 245, 260
  - Gaussian process, 397
  - individual-sequence, 199, 202, 273
  - linear model, 350, 363
  - LNML and linear model, 364
  - luckiness, *see* luckiness regret
  - metauniversal coding, 304
  - minimax, 182, 202, 208, 293
    - exponentiated, 216
    - unrestricted, 202
  - nonexponential families, 241
  - stochastic, 199
  - worst-case, 180
- regular histogram, 377
- regularity, 73, 103, 595
- relation, 82
- relations between divergences, 517
- relative
  - entropy, 103
  - redundancy, 265
- renormalized complexity, *see* RNML
- reparameterization, 67, 159
- reproducing kernel Hilbert space, 396
- residual sum of squares, 343
- resolvability, 483
- $\rho$ -divergence, 512
- ridge regression, 347
- Riemann zeta function, 243
- risk, 466, 515, 519, 528
  - cumulative, 467
  - empirical, 580
  - Hellinger, *see* Hellinger risk
  - in learning theory, 579
  - KL, *see* KL risk
- Rissanen lower bound, 454

- Rissanen renormalization, *see* RNML
- Rissanen, J., xxiv, 26
- RKHS, 396
- RNML, 306, 438, 439, 449
  - as LNML-1, 443
- road ahead, 592
- robust regression, 451
- robustness property, *see*
  - exponential family,
  - robustness property
- rotation, 277
- rule of succession, 258
- sacrifice, 320
- saddle point integration, 239
- safe, 535
- sample, 69, 72
  - empty, 53
  - size
    - unknown, 134
  - space, 6, 46
    - continuous, 46
    - discrete, 46
  - virtual, 258
- sanity check, 29, 413, 514, 525
- Sanov's theorem, 127, 636
- Sauer's lemma, 583
- score, 116
  - logarithmic, *see* loss,
  - logarithmic
- selection-of-variables, 25
- semiprequential, 135, 196, 379
- sender, 83, 172
- separation, 416
- sequence
  - gappy, 473
  - infinite, 53
  - random, 103
- sequential decomposition
  - property, *see* probability,
  - sequential decomposition
- set, 41
  - bounded, 44
  - compact, 44
- Shannon, C.E., 95
- Shannon-Fano code, *see* code, Shannon-Fano
- Shtarkov code, *see* NML
- significance level, 412, 421
- simplex, 42, 68
- simplicity, 29
- singular coding system, 80
- singularity
  - absolute, 507
  - mutual, 154
- slack function, *see* luckiness function
- SMML, *see* minimum message length, strict
- smooth, 65
- Solomonoff's approach, 8, 546
- Solomonoff, R.J., 8
- source symbol, 80
- space of observations, *see* sample space
- speech recognition, 437
- squared
  - error, *see* loss, squared
  - loss, *see* loss, squared
- SRM, 581
  - and PAC-Bayes, 587
- start-up problems, 198, 261, 430
- statistic, 284
- statistical learning theory, *see*
  - learning theory,
  - statistical

- statistical risk, *see* risk
- Statistician, 644
- statistics, 69, 72
  - Bayesian, 26, 73, 74, 531
  - objective, 546
  - subjective, 544, 573
  - frequentist, 73, 524, 591
  - nonparametric, 35
  - orthodox, *see* statistics, frequentist
  - sufficient, 286, 301, 441, 568, 603, 630
  - Kolmogorov, 571
  - traditional, *see* traditional statistics
- Stein's lemma, 221
- Stevens's model, *see* model, Stevens's
- Stirling's approximation, 127, 128
- stochastic
  - complexity, 412
  - extended, 578
  - universality
    - of Bayes, 244
- strong law of large numbers, *see* law of large numbers
- structural risk minimization, 581
- structure, 72, 416
  - function, 571
- subjective Bayes, 544, 573
- subjectivity, 160, 533, 548
- sufficient statistics, *see* statistics, sufficient
- sum of squares, 341
  - fitted, 344, 444
  - residual, 343, 450
- superefficiency, 455, 527
- supervised learning, 72
- support, 47
- support vector machine, 401, 583
- supremum, 44
- surface area, 218
- SVM, *see* support vector machine
- Sweden, 75
- symbol, 79
- Taylor expansion, *see* KL divergence, Taylor expansion
- terminology, 69
- test set, 72, 565
- theorem, 644
- $\Theta_0$ -sequence, 210
- time series, 269, 368, 508
  - computation, 431
- topology, 44
- tradeoff, 135, 411, 415, 581
- traditional statistics, 469, 482
- training set, 72, 73, 565
- transpose, 43
- triangle inequality, 480
- trigonometric functions, 337
- trivial model, 414
- true, 525
- two-part code, 174, 183
  - behavior, 142
  - computation, 430
  - conditional, 284
    - modified, 292
  - countable, 184
  - crude, 274
  - design, 152, 157
  - discretization, 272
  - incomplete, 274, 291
  - MDL, *see* two-part MDL
  - meta, *see* meta-two-part code
  - regret
    - asymptotic, 273

- sample size-dependent, 196, 272
- simplistic, 139
- simplistic, for Markov chains, 138
- stupid, 157
- two-part MDL, 132, 136, 476, 477, 512, 590
  - approximating, 150
  - code, *see* two-part code
  - computing, 150
  - consistency, 143, 153, 157, 502
  - consistency theorem, 478
  - convergence rate, 478
  - estimator, 645
  - for Markov chains, 133
  - parameter estimation, 485
- type, 127
- umbral calculus, 228
- underfitting, 143, 147
- uniform
  - central limit theorem, 289
  - code, *see* code, uniform
  - convergence, 232
  - convergence rate, 516
  - law of large numbers, 580
  - luckiness, 485
- uniformly universal model, 183
- union bound, 55, 164, 637
- unit simplex, 42, 68
- universal
  - code, *see* universal model
  - computer language, 8
  - Turing machine, 8
- universal code
  - computation, 428
- universal model, 172, 175, 178
  - adaptive, 193
  - Bayesian, 175, 183, 192, 231, 443, 452, 519
    - approximation, 197
    - Bernoulli, 233, 258, 285
    - boundary, 234
    - computation, 429
    - countable, 184
    - for normal family, 264
    - histogram, 378
    - minimax, 234
    - multinomial, 262
    - regret, 232
    - stochastic
      - regret/redundancy, 244
      - surprise, 472
  - Césaro, 474
  - conditional, *see* code, conditional
  - countable  $\mathcal{M}$ , 184
  - CUP, 370, 505, 509
    - Bayes, *see* meta-Bayes code, 379
    - redundancy, 384
    - two-part, *see* meta-two-part code, 379
  - finite, 178
  - Gaussian process, 396, 542
  - Liang-Barron, *see* code, conditional, Liang-Barron
  - linear regression, *see* linear regression, universal model
  - luckiness, *see* luckiness universal model
  - minimax, 180, 181
  - NML, *see* NML histogram, 378
  - nonuniform, 185, 186, 187

- optimal, 180, 181, 202
- plug-in, 197, 462
  - asymptotics, 260
  - computation, 430
  - for exponential families, 260
  - for regression, 448
  - model selection, 431
  - multinomial, 262
  - redundancy, 265
  - regret, 260, 265
  - start-up problem, 261
- prequential, 190, 194, 195, 419, 461, 493
- semiprequential, 196
- stochastic, 201
- top-level, 406
- two-part, *see* two-part code
- uniform, 187
- uniformly, **183**
- universality
  - almost sure, 199
  - expected, 199
  - individual-sequence, 199
  - stochastic, 199
- utility, 532, 535, 540
  
- variance, 48
- VC-dimension, 582
- vector, 43
- virtual data, 347
- volume, 453
  - and complexity, *see* complexity, and volume
  - relative to divergence, 222
- Vovk, V., 576
- Vulcan, 545
  
- Wallace, C., 556
- Wallace-Freeman estimator, 497, 560
- warning, 523
- wavelets, 337
- weak
  - frequentist paradigm, 526
  - prequential principle, 528, 563
- weather forecasting, 529, 572
- Webb, G., 30
- weighted averaging, 26
- worst case
  - vs. average case, 451
- worst-case codelength, *see* codelength, minimax
- WPP, *see* weak prequential principle
- wrong-yet-useful principle, 33
  
- zero prior problem, 545
- zero-sum game, 637
- zeta function, 243