Bibliography

Alon, N., Ben-David, S., Cesa-Bianchi, N. and Haussler, D. (1997). Scale sensitive dimensions, uniform convergence and learnability. *J. ACM* **44** 615–631.

Audibert, J.-Y. (2004a). Aggregated estimators and empirical complexity for least square regression. *Ann. Inst. H. Poincaré Probab. Statist.* **40** 685–736.

Audibert, J.-Y. (2004b). PAC-Bayesian statistical learning theory. Ph.D. thesis, Univ. Paris 6. Available at http://cermics.enpc.fr/ audibert/.

Barron, A. (1987). Are Bayes rules consistent in information? In *Open Problems in Communication and Computation* (T. M. Cover and B. Gopinath, eds.) 85–91. Springer, New York.

Barron, A. and Yang, Y. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564–1599.

Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection by penalization. *Probab. Theory Related Fields* **113** 301–413.

Blanchard, G. (1999). The "progressive mixture" estimator for regression trees. *Ann. Inst. H. Poincaré Probab. Statist.* **35** 793–820.

Blanchard, G. (2001). Mixture and aggregation of estimators for pattern recognition. Application to decision trees. Ph.D. thesis, Univ. Paris 13. Available at http://ida.first.fraunhofer.de/ blanchard/.

Blanchard, G. (2004). Un algorithme accéléré d'échantillonnage Bayésien pour le modèle CART. *Rev. Intell. Artificielle* **18** 383–410.

Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam* (D. Pollard, ed.) 55–87. Springer, New York.

Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves. *Bernoulli* **4** 329–375.

Birgé, L. and Massart, P. (2001a). A generalized C_p criterion for Gaussian model selection. Preprint. Available at http://www.math.u-psud.fr/ massart/.

Birgé, L. and Massart, P. (2001b). Gaussian model selection. *J. Eur. Math. Soc.* **3** 203–268.

Blum, A. and Langford, J. (2003). PAC-MDL bounds. *Computational Learning Theory and Kernel Machines. 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop. COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003, Proceedings. Lecture Notes in Comput. Sci.* **2777** 344–357. Springer, New York.

Catoni, O. (2002). Data compression and adaptive histograms. In *Foundations of Computational Mathematics. Proceedings of the Smalefest 2000* (F. Cucker and J. M. Rojas eds.) 35–60. World Scientific.

Catoni, O. (2003). Laplace transform estimates and deviation inequalities. *Ann. Inst. H. Poincaré Probab. Statist.* **39** 1–26.

Catoni, O. (2004). Statistical learning theory and stochastic optimization. *Ecole*

d'Été de Probabilités de Saint-Flour XXXI—2001. Lecture Notes in Math. **1851** 1–270. Springer, New York.

Cristianini, N. and Shawe Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel Based Learning Methods.* Cambridge Univ. Press.

Feder, M. and Merhav, N. (1996). Hierarchical universal coding. *IEEE Trans. Inform. Theory* **42** 1354–1364.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer, New York.

Langford, J. and McAllester, D. (2004). Computable shell decomposition bounds. *J. Machine Learning Research* **5** 529–547.

Langford, J. and Seeger, M. (2001a). Bounds for averaging classifiers. Technical report CMU-CS-01-102, Carnegie Mellon Univ. Available at http://www.cs.cmu.edu/ jcl.

Langford, J., Seeger, M. and Megiddo, N. (2001b). An improved predictive accuracy bound for averaging classifiers. *International Conference on Machine Learning* **18** 290–297.

Littlestone, N. and Warmuth, M. (1986). Relating data compression and learnability. Technical report, Univ. California, Santa Cruz. Available at http://www.soe.ucsc.edu/ manfred/pubs.html.

McAllester, D. A. (1998). Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (Madison, WI, 1998)* 230–234. ACM, New York.

McAllester, D. A. (1999). PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (Santa Cruz, CA, 1999)* 164–170. ACM, New York.

McDiarmid, C. (1998) Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics* (M. Habib, C. McDiarmid and B. Reed, eds.) 195–248. Springer, New York.

Mammen, E. and Tsybakov, A. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829.

Ryabko, B. Y. (1984). Twice-universal coding. *Problems Inform. Transmission* **20** 24–28.

Seeger, M. (2002). PAC-Bayesian generalization error bounds for Gaussian process classification. *J. Machine Learning Research* **3** 233–269.

Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C. and Anthony, M. (1998). Structural risk minimization over data-dependent hierarchies. *IEEE Trans. Inform. Theory* **44** 1926–1940.

Shawe-Taylor, J. and Cristianini, N. (2002). On the generalization of soft margin algorithms. *IEEE Trans. Information Theory* **48** 2721–2735.

Tsybakov, A. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166.

Tsybakov, A. and Van de Geer, S. (2005). Square root penalty: Adaptation to the margin in classification and in edge estimation. *Ann. Statist.* **33** 1203–1224.

Van de Geer, S. (2000). *Applications of Empirical Process Theory.* Cambridge Univ. Press.

Vapnik, V. N. (1998). *Statistical Learning Theory.* Wiley, New York.

Vert, J.-P. (2000). Double mixture and universal inference. Preprint. Available at http://cbio.ensmp.fr/ vert/publi/.

Vert, J.-P. (2001a). Adaptive context trees and text clustering. *IEEE Trans. Inform. Theory* **47** 1884–1901.

Vert, J.-P. (2001b). Text categorization using adaptive context trees. *Proceedings*

of the CICLing-2001 Conference (A. Gelbukh, ed.) 423–436. *Lecture Notes in Comput. Sci.* **2004**. Springer, New York.

Willems, F. M. J., Shtarkov, Y. M. and Tjalkens, T. J. (1995). The context-tree weighting method: Basic properties. *IEEE Trans. Inform. Theory* **41** 653–664.

Willems, F. M. J., Shtarkov, Y. M. and Tjalkens, T. J. (1996). Context weighting for general finite-context sources. *IEEE Trans. Inform. Theory* **42** 1514–1520.

Zhang, T. (2006a). From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *Ann. Statist.* **34** 2180–2210.

Zhang, T. (2006b). Information-theoretic upper and lower bounds for statistical estimation. *IEEE Trans. Inform. Theory* **52** 1307–1321.