

Cambridge University Press

978-0-521-11862-0 - Neural Network Learning: Theoretical Foundations

Martin Anthony and Peter L. Bartlett

Index

[More information](#)

Author index

- Aho, A. V., 306, **365**
 Akaïke, H., 227, **365**
 Albertini, F., 282, **365**
 Aldous, D., 28, **365**
 Alexander, K. S., 58, **365**
 Alon, N., 41, 183, 256, 267, **365**
 Anderson, J. A., 9, **365**
 Angluin, D., 27, 28, 340, 361, **365**
 Anthony, M., 9, 28, 58, 73, 139, 163, 183, 192, 217, 227, 268, 296, 330, 339, **365, 366, 377**
 Assouad, P., 41, **366**
 Auer, P., 339, 355, **366**
- Babai, L., 183, **366**
 Barron, A. R., 9, 216, 227, 240, 268, 282, 355, **366, 367**
 Bartlett, P. L., 28, 85, 129, 138, 139, 150, 183, 192, 216, 217, 227, 246, 256, 257, 267, 268, 282, 296, 329, 339, 340, 355, 356, **366–368, 372, 374–377**
 Barve, R. D., 28, **368**
 Baum, E. B., 85, 330, 340, 341, **368**
 Baxter, J., 72, 107, 217, 356, **368, 375**
 Ben-David, S., 28, 72, 107, 183, 256, 267, 339, 340, **365, 367, 368**
 Benedek, G. M., 28, **368**
 Benedetti, R., 129, **368**
 Bernstein, S. N., 363, **368**
 Biggs, N., 9, 58, 339, **366**
 Billingsley, P., 27, **369**
 Birgé, L., 267, **369**
 Bishop, C. M., 9, **369**
 Block, H. D., 330, **369**
 Blum, A., 28, 339, **369**
 Blumer, A., 27, 58, 306, 315, 329, **369**
 Bollobás, B., 41, **369**
 Boser, B. E., 139, 217, **369**
 Boulton, D. M., 227, **378**
- Breiman, L., 356, **369**
 Brent, R. P., 340, **369**
 Brightwell, G., 28, 330, **366**
 Buescher, K. L., 257, **369**
 Burges, C. J. C., 217, **377**
 Bylander, T., 330, **369**
- Carl, B., 216, **369**
 Cesa-Bianchi, N., 27, 183, 256, 267, **365, 368, 369**
 Chalasani, P., 28, **369**
 Chari, S., 41, **370**
 Chazan, D., 85, **371**
 Chernoff, H., 360, **370**
 Chervonenkis, A., 72, **378**
 Chervonenkis, A. Y., 27, 41, 58, 150, 192, 246, **378**
 Cormen, T., 315, **370**
 Cortes, C., 217, 355, 356, **370**
 Cover, T. M., 41, 85, **370**
 Cybenko, G., 10, **370**
- Darken, C., 355, **370**
 DasGupta, B., 130, 339, **370**
 Devroye, L., 27, 72, 106, 362, 363, **370**
 Donahue, M., 355, **370**
 Downs, T., 356, **371**
 Drucker, H., 355, 356, **370**
 Duda, R. O., 27, 138, **370**
 Dudley, R. M., 41, 58, 163, 216, 256, **370, 378**
- Ehrenfeucht, A., 27, 58, 72, 306, 315, 329, **369, 371**
 El-Jaroudi, A., 41, **375**
 Erlich, Y., 85, **371**
- Faragó, A., 227, 329, **371**
 Fefferman, C., 282, **371**
 Feller, W., 27, 358, **371**

- Fischer, P., 27, 28, **367, 369**
 Fomin, S. V., 282, 359, **373**
 Frankl, P., 41, 183, **366, 371**
 Frean, M., 340, 356, **371, 375**
 Freund, Y., 28, 217, 355, 356, **371, 376**
 Friedman, J., 356, **371**
- Gallant, A. R., 216, 240, 282, **375**
 Gallant, S., 329, **371**
 Garey, M. R., 315, 329, **371**
 Goldberg, P. W., 107, 129, **371**
 Goldman, S. A., 28, **371**
 Golea, M., 340, **372**
 Grenander, U., 227, **371**
 Grove, A. J., 356, **371**
 Guo, Y., 217, **372**
 Gurvits, L., 183, 216, 217, 355, **370, 372**
 Guyon, I. M., 139, 217, **369**
 Györfi, L., 27, 106, 362, 363, **370**
- Hagerup, T., 360, **372**
 Hancock, T. R., 340, **372**
 Hart, P. E., 27, 138, **370**
 Hastie, T., 356, **371**
 Haussler, D., 27, 41, 58, 72, 85, 163, 183, 216, 240, 246, 256, 267, 282, 296, 306, 315, 329, **365, 368, 369, 371, 372**
 Haykin, S., 9, **372**
 Hebb, D., 9, **372**
 Helmbold, D. P., 28, **372**
 Herbster, M., 339, **366**
 Hertz, J., 9, 227, **372**
 Hinton, G. E., 9, 130, **373, 376**
 Hoeffding, W., 361, **372**
 Holden, S. B., 163, **366**
 Horn, K. S. V., 329, **372**
 Hornik, K., 10, **372**
 Horváth, M., 192, **373**
 Höffgen, K.-U., 28, 329, **367, 372**
- Ishai, Y., 296, **366**
 Itai, A., 28, **368**
- Jackel, L. D., 355, **370**
 Jackson, J., 28, **373**
 Jacobs, R. A., 130, **373**
 Jerrum, M. R., 107, 129, **371**
 Ji, C., 85, **373**
 Johnson, D. S., 315, 329, **371, 373**
 Jones, L. K., 216, 340, **373**
 Jordan, M. I., 130, **373**
 Judd, J. S., 339, **373**
- Karmarkar, N., 329, **373**
 Karpinski, M., 107, 130, **373**
- Kearns, M. J., 9, 27, 28, 72, 139, 163, 227, 306, 340, **371–373**
 Khovanskii, A. G., 129, **373**
 Koiran, P., 85, 129, 130, 139, 216, 355, **372, 373**
 Kolmogorov, A. N., 150, 183, 282, 359, **373, 374**
 Kowalczyk, A., 41, **374**
 Krogh, A., 9, 227, **372**
 Krzyżak, A., 106, 130, 227, 268, **374**
 Kuh, A., 28, **374**
 Kulkarni, S. R., 28, 183, 257, **367**
 Kumar, P. R., 257, **369**
 Kwek, S., 355, **366**
- Laird, P., 27, **365**
 Lee, W. S., 85, 139, 216, 217, 282, 329, 355, 356, **374, 376**
 Leiserson, C., 315, **370**
 Leshno, M., 10, **374**
 Levy, A., 85, **371**
 LeCun, Y., 355, **370**
 Li, M., 27, 28, **373, 374**
 Lin, V., 10, **374**
 Lindenbaum, M., 72, 107, **368**
 Linder, T., 106, 130, 227, 268, **374**
 Linial, N., 227, **374**
 Littlestone, N., 27, 306, **372**
 Long, P. M., 28, 58, 107, 183, 246, 256, 267, 268, **367, 368, 372, 374**
 Lorentz, G. G., 163, **374**
 Lugosi, G., 27, 58, 72, 106, 130, 192, 227, 268, 329, 362, 363, **368, 370, 371, 373–375**
- Maass, W., 27, 85, 355, **366, 375**
 Macintyre, A. J., 106, 107, 130, **373, 375**
 Maillot, V., 282, **365**
 Maiorov, V., 129, **368**
 Makhoul, J., 41, **375**
 Makovoz, Y., 216, **375**
 Mansour, Y., 28, 227, **371, 373, 374**
 Marchand, M., 340, **372**
 Mason, L., 217, 356, **375**
 Massart, P., 267, **369**
 Matousek, J., 41, **375**
 McCaffrey, D. F., 216, 240, 282, **375**
 Meir, R., 129, **368**
 Mhaskar, H., 10, **375**
 Milnor, J., 107, **375**
 Minsky, M., 9, **375**
 Miyano, S., 28, **377**
 Muroga, S., 58, 330, **375**
- Natarajan, B. K., 9, 183, 246, 296, 315, **375**

- Ng, A. Y., 227, **373**
 Nilsson, N. J., 85, 330, **376**
 Nowlan, S. J., 130, **373**
- Palmer, R. G., 9, 227, **372**
 Papert, S., 9, **375**
 Patel, J. K., 364, **376**
 Petrack, S., 85, **371**
 Petsche, T., 28, **374**
 Pinkus, A., 10, **374**
 Pisier, G., 216, **376**
 Pitt, L., 315, 330, 339, **376**
 Pollard, D., 58, 150, 163, 192, 246, 256,
 267, 268, 282, **376**
 Posner, S. E., 183, 256, **367**
 Powell, M. J. D., 130, **376**
 Preparata, F. P., 329, **373**
 Psaltis, D., 85, **373**
- Quinlan, J. R., 356, **376**
- Read, C. B., 364, **376**
 Ripley, B. D., 9, **376**
 Risler, J.-J., 129, **368**
 Rissanen, J., 227, **376**
 Rivest, R. L., 28, 227, 315, 339, **369**,
 370, **374**
 Rohatgi, P., 41, **370**
 Ron, D., 227, **373**
 Rosenblatt, F., 9, 330, **376**
 Rosenfeld, E., 9, **365**
 Rub, C., 360, **372**
 Rumelhart, D. E., 9, **376**
- Sakurai, A., 85, 129, **376**
 Sauer, N., 41, **376**
 Schapire, R. E., 27, 139, 163, 217, 355,
 356, 371, **373**, **376**
 Schläfli, L., 41, **377**
 Schocken, S., 10, **374**
 Schuurmans, D., 73, 217, 356, 371, **377**
 Schwartz, R., 41, **375**
 Schölkopf, B., 217, **377**, **378**
 Sellie, L. M., 27, **373**
 Shamir, E., 27, **369**
 Shawe-Taylor, J., 28, 58, 73, 139, 217,
 227, 296, 330, **366**, **368**, **372**, **377**
 Shelah, S., 41, **377**
 Shinohara, A., 28, **377**
 Siegelmann, H. T., 339, **370**
 Simon, H. U., 27, 72, 164, 192, 268,
 329, **369**, **372**, **377**
 Sloan, R. H., 27, **377**
 Slud, E., 363, **377**
 Smola, A. J., 217, **377**, **378**
 Sontag, E. D., 41, 85, 106, 129, 130,
 139, 282, 339, 340, 355, **365**, **370**,
373, **375**, **377**
- Srinivasan, A., 41, **370**
 Steele, J. M., 41, **377**
 Sternberg, S., 364, **377**
 Stinchcombe, M., 10, **372**
 Stone, C., 27, **377**
 Sussmann, H. J., 139, 282, 339, **377**
- Talagrand, M., 58, **378**
 Tate, R. F., 364, **378**
 Thomas, J., 85, **370**
 Tibshirani, R., 356, **371**
 Tihomirov, V. M., 150, 183, **374**
 Tikhomirov, V. M., 163, **378**
 Tomkins, A., 28, **373**
- Ullman, J. D., 306, **365**
- Valiant, L. G., 27, 72, 306, 315, 330,
 339, 361, **365**, **371**, **376**, **378**
 van der Vaart, A. W., 216, **378**
 Vapnik, V., 72, **378**
 Vapnik, V. N., 9, 27, 41, 58, 73, 139,
 150, 192, 217, 227, 240, 246, 296,
 355, **369**, **370**, **378**
 Vazirani, U., 9, 28, 340, **365**, **373**
 Vidyasagar, M., 9, 216, 246, **378**
 Vitányi, P. M. B., 28, **374**
 Vu, V. H., 340, **378**
- Wahba, G., 217, **378**
 Wallace, C. S., 227, **378**
 Warmuth, M. K., 27, 58, 306, 315, 329,
 339, 355, **366**, **369**, **372**
 Warren, H. E., 107, **378**
 Watkins, D. S., 329, **378**
 Wellner, J. A., 216, **378**
 Welzl, E., 41, **372**
 Wenocur, R. S., 41, **378**
 White, H., 10, 216, **372**, **378**
 Williams, R. J., 9, **376**
 Williamson, R. C., 85, 129, 139, 216,
 217, 227, 246, 267, 268, 282, 296,
 329, 355, **367**, **368**, **372**, **374**,
377, **378**
- Zeger, K., 227, 268, **375**
 Zuev, Y. A., 58, **378**

Subject index

- (inner product), 3
- $[n]$ (the set $\{1, 2, \dots, n\}$), 156
- $\lceil \cdot \rceil$ (ceiling function), 56
- $\lfloor \cdot \rfloor$ (floor function), 56
- \otimes (Kronecker product), 172
- $\mathcal{A}(z, \epsilon)$ (approximate-SEM algorithm), 237
- absolute loss, 267, 284
 - average over multiple outputs, 288
- accuracy parameter, 16
- activation function, 5, 76
 - satisfying a Lipschitz condition, 199
- Adaboost, 352, 355
 - and weighted sample error minimization, 353
 - learning in the restricted real classification model, 355
- adaptivity, 226
- affine subspace, 3
- agnostic pac learning, 27
- Akaike's information criterion, 227
- algorithmics of supervised learning, 9, 299–356
- analytic functions, 35
- APPROX- F -FIT decision problem, 314
- approximate interpolation, 289, 296
 - generalization from, 291
 - and fat-shattering dimension, 294
 - and pseudo-dimension, 293
 - strong generalization from, 290
 - and band dimension, 292
 - and pseudo-dimension, 292
- approximate-SEM algorithm, 236, 237, 258
 - efficient, 304
 - efficient randomized, 309
 - for a class of vector-valued functions, 288
 - for a graded function class, 302
 - for convex combinations, 342, 346
 - sample complexity, 258
- approximation error of a function class, 15
- approximation issues in supervised learning, 1, 9, 18
- artificial neural networks, *see* neural networks
- asymptotic optimality of pattern classification techniques, 17
- B (bound on $|\hat{y} - y|$ in real prediction problem), 233
- band(F, η) (band dimension), 292
- band dimension, 292
 - and pseudo-dimension, 292
- Bayes optimal classifier, 17
- Bernoulli random variable
 - estimating the probability of, 59, 273
- Bernstein's inequality, 278, 283, 363
- B_F (subgraph class), 153, 154
- B_f (indicator function of region below graph of f), 153
- binary classification, 8, 13
 - restricted model, 23, 52, 263, 289
 - and existence of weak predictors, 351
 - and probably approximately correct model, 306
 - efficient learning, 305, 306, 315
 - estimation error, 53, 58
 - sample complexity, 52, 58
 - sample complexity, 184–192
- binary input, 4
- binomial distribution, 19
- binomial theorem, 358
- boolean perceptron, 316–330

- efficient consistent-hypothesis-finder, 323
 - sample complexity, 52
- boosting algorithms, 355
- bounded variation functions, 159
 - covering numbers, 180, 183
 - fat-shattering dimension, 160, 163
- BP , BP_n (functions computed by the boolean perceptron), 316
- consistent-hypothesis-finders, 322
 - not efficiently learnable, 319
- BP -FIT
 - decision problem, 316
 - reduction to VERTEX COVER, 317
- Cauchy-Schwarz inequality, 359
- $CC(\cdot)$ (number of connected components), 30
- chaining, 55–58, 265, 267
- Chebyshev inequality, 360
- Chernoff bounds, 360
- classification, 14
- classification learning, *see* binary classification, real classification
- closed under addition of constants, 188
- closed under scalar multiplication, 162
- closure convexity, 269
 - and sample complexity, 271, 277
 - and variances, 278
 - definition, 271
- closure of a function class, 270
- $co(S)$ (convex hull of S), 204
- coin toss, 19
- compactness, 271, 282
- complexity
 - automatic choice of, 219–227
 - of a network, 8
 - penalty, 8, 221, 222
 - regularization, 227
 - theory, 312, 315
- computability, 299
- computation
 - bit cost model of, 301
 - issues in supervised learning, 9
 - units, 74
- computational complexity, 2, 299, 306, 312
- conditional expectation
 - approximating, 232
 - best approximation, 278
 - empirical, 346
- confidence parameter, 16
- connected components, 30
- connections, 74
- consistent hypotheses, 305
- consistent learning algorithms, 24, 27
- consistent-hypothesis-finders, 52, 315
 - efficient, 305
- Construct** (learning algorithm for convex combinations), 346, 355
 - sample complexity, 348
- constructive learning algorithms, 340, 342–356
- continuous from the right, 195
- convex class
 - sample complexity, 263
- convex classes, 269–283
 - sample complexity, 277
 - slow convergence with absolute loss, 296
- convex combinations
 - approximation rate, 203, 216
 - constructive approximation, 342, 355
 - constructive learning algorithms, 342–351
 - covering number bounds, 205
 - large margin SEM algorithm, 352
- convex function
 - Jensen's inequality, 358
- convex hull, 204
- covering numbers
 - and dimensions, 165–183
 - and uniform convergence, 140–150
 - as generalization of growth function, 241
 - bounds, 247–257
 - pseudo-dimension versus fat-shattering dimension, 181
 - bounds in terms of fat-shattering dimension, 175, 248
 - d_1 , 241, 247
 - bounds in terms of pseudo-dimension, 251
 - d_∞ , 241, 247, 268
 - effects of composition with a Lipschitz function, 206
 - effects of scaling, 206
 - lower bound in terms of fat-shattering dimension, 178
 - of composition of function classes, 197
 - of the loss function class, 242
 - relationship with packing numbers, 166, 183
- critical point, 364
- critical value, 364
- d_1 covering numbers, *see* covering numbers, d_1
- d_1 packing numbers, *see* packing numbers, d_1
- d_∞ covering numbers, *see* covering numbers, d_∞
- $d_{L_1(P)}$ ($L_1(P)$ pseudometric), 251

- packing numbers, *see* packing numbers, $d_{L_1(P)}$
- $d_{L_2(P)}$ ($L_2(P)$ metric), 255
- d_{L_∞} (L_∞ metric), 197, 236
- $d_\infty^p(\cdot, \cdot)$ (distance between vectors of elements of a metric space), 197
- data generation model, 231
- decision boundary, 3, 4
- decision problem, 313
- decision rule, 60
 - randomized, 273
- decision-theoretic learning models, 296
- δ (confidence parameter), 16
- dichotomies
 - counting in parameter space, 30
- $\dim(\cdot)$ (linear dimension), 37
- distribution
 - changing, 28
- distribution-independence, 18
- early stopping, 225
- efficient learning, 299–306
 - characterization, 312
- efficient learning algorithm, 306
 - and fat-shattering dimension, 303
 - and H -FIT, 313
 - and NP-hardness of the H -CONSISTENCY decision problem, 314
 - and NP-hardness of the H -FIT decision problem, 314
 - and VC-dimension, 303
- definition, 302
- efficient randomized SEM algorithm is necessary, 311
- efficient randomized SEM algorithm is sufficient, 309
- empirical cover, 255
- empirical error, 234
- entropy, 79, 85
- ϵ (accuracy parameter), 16
- ϵ -good hypothesis, 16
- ϵ -packing, 165
- ϵ -separated, 55, 165
- $\epsilon_0(m, \delta)$ (estimation error), 17
- $\epsilon_L(m, \delta, B)$ (estimation error), 234
- $\epsilon_L(m, \delta, \gamma)$ (estimation error), 137
- $er_P^\gamma(\cdot)$ (error of a real-valued function with respect to P and γ), 136
- $\hat{er}_z^\gamma(\cdot)$ (sample error of a real-valued function with respect to γ), 184
- $\hat{er}_z^\ell(\cdot)$ (ℓ -sample error of a real-valued function), 285
- $\hat{er}_z(\cdot)$ (sample error of a binary-valued function), 15
- $\hat{er}_z(\cdot)$ (sample error of a real-valued function), 234
- $er_P^\ell(\cdot)$ (ℓ -error of a real-valued function), 284
- $er_\mu(\cdot, t)$ (error of a binary-valued function with respect to target t), 24
- $er_P(\cdot)$ (error of a binary-valued function), 15
- $er_P(\cdot)$ (error of a real-valued function), 233
- error
 - and error estimate, 8
 - convergence rate, 17, 27
 - expected value of, 26
 - relationship with the restricted model, 27
 - in restricted model of binary classification, 24
 - of a binary-valued function, 15
 - of a real-valued function, 232
- estimation error
 - convergence rate, 18
 - definition, 17, 18, 137, 234
 - for classes with finite VC-dimension, 43
 - for finite classes, 21, 235
 - inherent, 18
- estimation issues in supervised learning, 2, 4, 9, 18
- Euler's inequality, 358
- \bar{F} (closure of F), 270
- F_n (function class with complexity parameter n), 300
- f^* (conditional expectation), 279
- f_a (best approximation to conditional expectation), 278
- factorial
 - Stirling's approximation, 358
- fat-shattering dimension, 8, 159–163
 - and covering numbers, 174, 180, 182, 183
 - and learnability of real function classes, 258
 - and packing numbers, 174
 - arbitrary rate, 182
 - characterization of learnability, 262–267
 - definition, 159
 - finite, 159
 - but pseudo-dimension infinite, 162
 - of parameterized classes
 - with bounded number of parameters, 196
 - with bounded parameters, 203
 - relationship with pseudo-dimension, 162, 163
- $\text{fat}_F(\gamma)$ (fat-shattering dimension), 159

- feed-forward networks, 74
- finite function classes, 19–22, 234–236
- fully connected between adjacent layers, 75
- γ -dimension, 159
- γ -shattered, 159
- Gaussian elimination, 329
- GE(p, m, k) (Chernoff bound), 361
- general position, 30
 - and linear independence, 32
- generalization, 5
- graded function class, 299
- gradient descent, 6, 9, 225, 339
- graph colouring, 332
- graph theory, 316
- growth function, 29–35
 - and VC-dimension, 39
- H (binary-valued function class), 15
- H^k (functions computed by simple perceptrons with fan-in no more than k), 321
- H_n (binary function class with complexity parameter n), 300
- H -CONSISTENCY decision problem, 313
- H -FIT decision problem, 313
- Hamming distance, 80, 178
- Hilbert space, 271, 282
- Hoeffding's inequality, 361
 - bounds on tails of binomial distribution, 20
 - in uniform convergence proof, 50, 55
- Hölder inequalities, 359
- hyperplanes, 31
- hypothesis, 15
 - representation, 299
- $\Im(z)$ (imaginary part of z), 275
- independence of training examples, 27
- inner product, 3, 343
- input, 14
 - space, 7
 - units, 74
 - vector, 2
 - weights, 5
- interior point methods, 214
 - for linear programming, 323
- interpolation, 296
- interpolation models, 289–295
- Jacobian, 364
- Jensen's inequality, 358
- k -COLOURING decision problem, 332
- k -plane, 33
- Karmarkar's algorithm, 323
- Kronecker product, 172
- L (learning algorithm), 16
- $L_1(P)$ pseudometric, 251
- $L_2(P)$ covering numbers, 255
- L_∞ covering numbers
 - sample complexity bounds in terms of, 237
- L_∞ metric, 196, 236
- ℓ_f (loss function for f), 233
- ℓ_F (loss class), 284
- ℓ^n (loss function for approximate interpolation), 291
- ℓ^s (loss function for s outputs), 286
 - covering numbers, 287
- ℓ -error, 284
- ℓ -layer network, 75
- ℓ -sample error, 285
- labelled examples, 6, 14
- labels, 1, 14
 - noisy, 27
- large margin classification, 8, 135
 - and generalization from approximate interpolation, 294
- law of large numbers, 19
- layers, 74
- LE(p, m, k) (Chernoff bound), 361
- learnability, 17, 234
 - efficient, 308
- learner, 13
- learning
 - as optimization, 307–315
 - pattern classification, *see* binary classification, real classification
 - real-valued functions, *see* real prediction
- learning algorithms
 - based on approximate-SEM algorithm, 259
 - definition, 13, 16, 136, 233
 - enumerative, 329
- linear computation units
 - fat-shattering dimension, 213
 - pseudo-dimension, 155
- linear programming
 - efficient algorithms, 323
 - for boolean perceptron learning algorithms, 322
 - for BP_n learning algorithms, 329
 - for classification learning with linear functions, 329
- linear subspace, 33
- linear threshold networks, 76
 - approximate sample error minimization, 340
 - feed-forward, 74
 - growth function, 77

- hardness of learning, 331, 335, 338, 339
- simulation with sigmoid network, 84, 85
- VC-dimension, 74–85
 - lower bounds, 80, 82, 85
 - upper bounds, 77, 85
- with bounded fan-in
 - efficient learning algorithm, 350
- Lipschitz condition, 199
- local minima, 6, 339
- logarithm
 - inequalities, 357
- loss class, 284
 - covering numbers, 285
- loss functions, 233, 284–289, 296
 - bounded, 284
 - satisfying a Lipschitz condition, 245, 285
- $m_0(\epsilon, \delta)$ (sample complexity), 16
- $m_L(\epsilon, \delta)$ (sample complexity), 234
- $m_L(\epsilon, \delta, \gamma)$ (sample complexity), 137
- $\mathcal{M}(\epsilon, W, d)$ (ϵ -packing number of W with respect to d), 165
- $\mathcal{M}_1(\epsilon, H, k)$ (uniform packing number with respect to d_1), 165
- $\mathcal{M}_2(\epsilon, H, k)$ (uniform packing number with respect to d_2), 165
- margins, 291, 351
- Markov's inequality, 360
- matrix
 - determinant, 35
 - row-rank, 37
- measurability conditions, 15
- measure
 - Lebesgue, 35, 364
 - outer, 364
 - theory, 27
- method of sieves, 227
- minimum description length principle, 226, 227
- minimum message length, 227
- misclassification probability, 8
- model selection, 218–227
- monomials, 171
- μ (probability distribution), 24
- multi-layer networks, *see* neural networks, multi-layer
- multisets, 156
- $N_{\sigma}^2, N_{\sigma, n}^2$ (two-layer sigmoid networks with two first-layer units), 337
- N_{σ}^2 -APPROX-SEM problem, 337
- N_n^k, N_n^k (two-layer linear threshold networks with k first-layer units), 335
 - and graph colouring, 335
- $N_{\wedge}^k, N_{\wedge, n}^k$ (conjunctions of k linear threshold units on binary inputs), 332
 - and graph colouring, 333, 339
- N_{\wedge}^k -CONSISTENCY decision problem, 332
- NP-hardness, 335
- $N_{\sigma}^k, N_{\sigma, n}^k$ (two-layer sigmoid networks with k first-layer units), 338
- N_{σ}^p -APPROX-SEM problem, 338
- net input, 84
- neural networks
 - architecture, 74
 - biological, 9
 - classes with infinite pseudo-dimension, 265
 - dimensions, 193–217
 - general, 7
 - hardness of learning, 331–341
 - Lipschitz in parameters
 - covering number bounds, 199
 - fat-shattering dimension bounds, 202
 - multi-layer, 74
 - sample complexity bounds, 261
 - state, 13, 76, 299
 - two-layer
 - approximate-SEM algorithm, 350
 - as convex combinations of functions, 342
 - VC-dimension, 108–130
 - with bounded output weights as convex combination, 277
 - with bounded parameters
 - covering number bounds, 207–212
 - fat-shattering dimension bounds, 212–213
 - sample complexity, 262, 349
 - with finite weight set
 - sample complexity, 236
 - with multiple outputs, 286–289
 - sample complexity, 288, 296
 - with piecewise-linear activation functions
 - hardness of learning, 339
 - with piecewise-polynomial activation functions
 - sample complexity, 261
 - with real-valued output, 231
- noise, 14, 231, 266, 268
- non-convex classes
 - sample complexity, 263, 270
- normal distribution
 - tail bounds, 364
- norms
 - dual, 359

Subject index

387

- induced by probability distributions, 270
- NP-hardness, 312
- $\text{opt}_P(F)$ (approximation error of the class F of real functions), 233
- $\text{opt}_P(H)$ (approximation error of the class H of binary-valued functions), 15
- $\text{opt}_P^\gamma(F)$ (optimal large margin error of the class F), 185
- orthants of \mathbb{R}^m , 153
- output
 - space, 7
 - unit, 75
 - weights, 5
- P (probability distribution), 14
- PAC learning, *see* probably approximately correct learning
- packing, 80
- packing numbers, 55, 165–167
 - and quantized classes, 168
 - bounds in terms of fat-shattering dimension, 174
- d_1
 - bound in terms of fat-shattering dimension, 247
- $d_{L_1}(P)$
 - bounds in terms of pseudo-dimension, 251
 - uniform, 165
- parameter space
 - counting cells, 30, 32, 41
- parameterization
 - uniqueness of, 280
- pattern classification, 13
 - with binary-output networks, 11–130
 - with real-output networks, 131–227
- patterns, 1
- $\text{Pdim}(\cdot)$ (pseudo-dimension), 153
- perceptron, 2, 9, 22, 74, 77
 - binary-weight, 23
 - sample complexity, 23
 - convergence theorem, 323, 330
 - and real classification, 328
 - enumerative learning algorithm, 319
 - estimation error, 51
 - fan-in, 319
 - functions computable by, 7
 - k -bit, 23
 - sample complexity, 23
 - learning algorithm, 3, 9, 323, 329
 - and classification noise, 330
 - is not efficient, 328, 330
 - learning in the restricted model, 322–328
 - representational capabilities, 4
 - sample complexity, 51
 - shattering and affine independence, 36, 41
 - with binary inputs, 316
 - permutations on a double sample, 47, 242
 - swapping group, 58
 - symmetric group, 58
 - Φ -dimension, 170, 183
 - $\Phi\text{dim}(\cdot)$ (Φ -dimension), 170, 182
 - pigeonhole principle, 166
 - $\Pi_H(\cdot)$ (growth function), 29
 - P^m (product probability distribution), 15
 - polynomial transformation, 156, 163
 - pseudo-dimension, 156
 - polynomial-time algorithm, 313
 - predicting a real-valued quantity, *see* learning real-valued functions
 - probabilistic concepts, 192
 - and fat-shattering dimension, 163
 - probability distribution, 14
 - product, 15
 - probability estimation, 231
 - probability theory, 13, 27
 - probably approximately correct learning, 27, 306
 - and weak learning, 355
 - pseudo-dimension, 151–159, 163
 - and compositions with non-decreasing functions, 153
 - and d_∞ -covering numbers, 167, 183
 - and linear dimension, 154
 - from VC-dimension bounds, 194, 216
 - infinite, 265, 267
 - pseudo-shattered, 152
 - versus shattered, 152
 - $Q_\alpha(\cdot)$ (quantization operator), 167
 - $Q_\alpha(F)$ (quantized versions of functions in F), 248
 - quadratic loss, 231, 245, 267, 278, 284
 - average over multiple outputs, 286, 288
 - quadratic optimization, 214, 217
 - quantization, 167, 248
 - queries, 28, 340
 - radial basis function networks, 213
 - random number generator, 307
 - randomized algorithm, 307
 - polynomial-time, 313
 - rank
 - of a linear system of equations, 38
 - of a matrix, 37
 - real classification, 8, 131–227

- real labels
 - encoding information in, 265
 - noisy, 266
 - quantized, 266, 268
- real prediction, 8, 231–296
 - restricted model, 265, 268, 281
 - sample complexity, 258–268
- regularization
 - complexity, 227
 - weight decay, 225
- $R_L(m, n)$ (worst-case running time), 301
- RP (decision problems solvable in polynomial time using a randomized algorithm), 313
- sample complexity, 17
 - bounding with pseudo-dimension, 260
 - definition, 18, 137, 234
 - for a class of vector-valued functions, 288
 - gap between upper and lower bounds, 262, 269
 - inherent, 18
 - lower bounds in terms of
 - fat-shattering dimension, 188, 262
 - lower bounds in terms of VC-dimension, 59–73
 - of a closure convex class, 269
 - of a finite binary-valued class, 21, 235
 - restricted model, 25
 - upper bounds in terms of
 - fat-shattering dimension, 265
 - upper bounds in terms of pseudo-dimension, 265
 - upper bounds in terms of VC-dimension, 42–58
- sample error
 - as estimate of error, 19
 - definition, 15, 234
 - weighted, 352
- sample error minimization algorithms, 19, 42
 - efficient, 304
 - efficient randomized, 309
 - for a graded function class, 302
 - for finite class of real-valued functions, 235
 - estimation error, 235
- large margin, 184
 - estimation error, 187
 - sample complexity, 187
 - sample complexity bound involving pseudo-dimension, 191
- sample complexity, 54
 - weighted, 352
- Sard's Theorem, 364
- Sauer's Lemma, 41
 - generalization involving pseudo-dimension, 170, 183
 - linear algebraic proof, 183
- scalar product, 204, 271
- scale-sensitive dimension, 159, 291
- SEM algorithm, *see* sample error minimization algorithms
- SET-SPLITTING decision problem, 339
- $\text{sgn}(\cdot)$ (threshold function), 3, 76
- shattering, 35, 41, 151
 - width of, 159
 - witness of, 152, 159
- $\sigma(\cdot)$ (standard sigmoid function), 83
- sigmoid functions
 - linear independence, 275
 - standard, 5, 83
- sigmoid networks, 83
 - approximation, estimation, computation properties, 6
 - hardness of learning, 337–338, 340
 - invariances, 281
 - non-convexity, 275
 - two-layer, 5
 - uniqueness of parameterization, 280, 282
 - VC-dimension lower bounds, 83
 - VC-dimension upper bounds, 122–128
- sigmoid unit, 83
- simple perceptron, *see* perceptron
- Slud's inequality, 363
- $\text{span}(\cdot)$, 171
- spanning set of a vector space, 171
- Splitting (enumerative algorithm for simple perceptrons), 319, 329, 349
- squared error, *see* quadratic loss
- squashing function, 5
- states, 7
- step function, 76
 - versus sigmoid function, 83
- structural risk minimization, 227
- subgraph class, 153
- supervised learning, 1
 - applications, 1
 - definition, 13
- support vector machines, 217
- symmetrization, 46, 242, 285
- tails of distributions, 360
- target function, 23
- testing sample, 46
- θ (threshold of simple perceptron), 2
- thresholds, 3, 76
- time complexity, 299
- topology of function classes, 270, 282
- total variation, 160

Subject index

389

- touchstone class, 25, 27
 - and computational complexity, 340
 - and sample complexity, 240
 - of real functions, 240
- trace number of a set system, 41
- training, 1
 - data, 14
 - samples, 14, 24
- uniform convergence
 - and fat-shattering dimension, 267
 - for classes with finite VC-dimension, 43, 53
 - for finite classes, 19
 - for real functions, 241–246
 - rate, 266, 282
 - improved using variance information, 277
 - restricted model with zero mean noise, 281, 282
 - relative results, 71–72, 191, 266
 - with general loss functions, 285
- uniform distance, 196
- union bound, 21
 - in uniform convergence proof, 50, 55
- universal approximation, 6, 10
- Vapnik-Chervonenkis dimension, 8, 35–41
 - lower bounds for smoothly parameterized function classes, 85
 - of a vector space of real functions, 37
- VC-dimension, *see* Vapnik-Chervonenkis dimension
- VC-major classes, 163
- VC-number, 41
- vector space
 - basis, 37
 - dimension, 37
 - pseudo-dimension, 154
 - spanning set, 171
- VERTEX COVER decision problem, 317
- vertex cover of a graph, 316
- W (number of network parameters), 76
- w (weights of simple perceptron), 2
- weak binary classification learning, 355
- weight decay, 225
- weights, 76
 - of a simple perceptron, 2
- worst-case running time, 301
- X (input space), 7
- x (input vector), 2, 14
- X_n (input space with complexity parameter n), 300
- Y (output space), 7
- y (label), 14
- Z ($= X \times Y$), 14
- z (training sample), 14
- Z_n ($= X_n \times Y$), 300