Subject and Author Index

Note: Numbers, symbols, Greek letters, etc. are alphabetized as if spelled out in words. Page listings tagged with “n” refer to footnotes; those tagged with “x” refer to exercises.

Abe, Naoki, 134, 333
abort, 437
abstaining weak hypotheses, 278–279
algorithmic speed-ups using sparse, 279–281
domain-partitioning, 293–294
rule as, 287
accuracy, 3
active learning, 129–132
AdaBoost, 5–7
adaptiveness of, 10–11, 56
analysis of error (see error analysis of AdaBoost)
Bayes error not reached, 398–404
benchmark experiments, 11
and boost-by-majority, 448–452
and Chernoff bounds, 60–62, 448
confidence-rated (see confidence-rated AdaBoost)
convergence of, 243–251
as coordinate descent, 180–181
dynamics of, 239, 263–264x
effect on margins, 111–114
and estimating conditional probabilities, 202
exhaustive, 185
and exponential loss, 177–179
for face detection, 66–70
and filtering of examples, 88
and functional gradient descent, 190–191
initial distribution modified in, 54
as iterative projection algorithm, 232–237, 239–242
as limit of BrownBoost, 476–483
and logistic regression, 197–200, 252–255
loss compared to RankBoost, 348–351
and maximum entropy, 234
multiclass (see multiclass boosting)
and noise, 404–405, 483–484
and optimal risk, 385–387
and overfitting, 15–16, 99
pronunciation, 11n
sample run, 7–10
training error of, 54–56
trajectory of, 209
and universal consistency, 386–387
and vMW, 172x
AdaBoost.L, 197–200
convergence of, 265–266x
experiments, 484–485
versus LogitBoost, 223–224x
with prior knowledge, 213
AdaBoost.MH, 312–314
versus AdaBoost.MR, 363
Hamming loss of, 313
one-error of, 314–315
training error of, 315–316, 324
weak learner for, 313–314
AdaBoost.Mk, 334x
AdaBoost.MO, 322
experiments, 332–333
generalization error of, 334–337x
generalized, 327–328
and RankBoost, 369–370x
training error of, 323–325, 328–332
AdaBoost.M1, 305
experiments, 308–309
training error of, 306–307
weak learning assumption, 305–306
AdaBoost.MR, 361
versus AdaBoost.MH, 363
for multiclass logistic regression, 372–373x
one-error of, 361–363
training error of, 363
AdaBoost.M2, 369
AdaBoost.M, 115, 135x
AdaBoost.Mk, 133
AdaBoost.S, 408–410x
adaptiveness, 10–11, 56, 459
additive modeling, 219
ADTs. See alternating decision trees
adversarial player, 143, 151
affine threshold functions, 89x
Agarwal, Shivani, 369
all-pairs reduction, 325
and loss-based decoding, 337–338x
and RankBoost, 369–370x
training error of, 332
Allwein, Erin L., 333
almost sure convergence, 386
alphabet, 231
α-Boost, 162
and margin maximization, 162–163
and MW, 162
and regularization, 207–209
trajectory of, 205–207
See also NonAdaBoost
αt, choosing, 275–279
in RankBoost, 347–348
alternating decision trees, 290–296
boosting algorithm for, 293–294
defined, 291–292
interpretability, 294–296
and overfitting, 294
as sum of branch predictors, 292–293
ambiguity of natural languages, 364
amortized analysis, 147
analysis of AdaBoost’s error. See error
analysis of AdaBoost
analysis of generalization error. See error
bounds; generalization error
Anthony, Martin, 89
AnyBoost, 190
See also functional gradient descent
AP headlines dataset, 213–215
and confidence-rated predictions, 286
approximability of Boolean functions, 73x
approximate maximin strategy, 153
and boosting, 161
approximate minmax strategy, 152
and boosting, 161
arc-gv, 115–116
arg max, 493
arg min, 493
ASSEMBLE.AdaBoost, 218
Associated Press. See AP headlines dataset
AT&T, 316
Atlas, Les, 133
attributes. See features
axis-aligned rectangles
learning, 50x
linear separability of, 58–59
VC-dimension of, 50x
Azuma’s lemma, 61
bagging, 118
and margins theory, 120
and variance reduction, 118
Bakiri, Ghioum, 333
Bartlett, Peter L., 89, 132–133, 406
base classifiers. See weak hypotheses
base functions, 256–258
base hypotheses. See weak hypotheses
base learning algorithm. See weak learning algorithms
batch learning, 153–154
Baum, Eric B., 89
Baxter, Jonathan, 133
Bayes error, 377, 379
approached by AdaBoost, 386–387
and generalized risk, 407–408x
not achieved by AdaBoost, 398–404
not achieved by general risk minimization, 412–413x
and optimal risk, 380–382
Bayes optimal classifier, 379
and optimal predictor, 380
Bayes optimal error. See Bayes error
BBM. See boost-by-majority
behaviors. See dichotomies
benchmark datasets
experiments with AdaBoost, 11
experiments with added noise, 484–485
with multiclass boosting, 308–309
Bennett, Kristin P., 170, 220
Beygelzimer, Alina, 333
bias. See bias-variance analysis
bias-variance analysis, 117–122
and bagging, 118
and boosting, 118–120
definitions, 120
Bickel, Peter J., 406
big Oh, 492
binary relative entropy, 114–115, 232
basic properties, 135x
on vectors, 253
binomial coefficient, 498
binomial distribution, 498
and Hoeffding’s inequality, 30–31
lower bound on, 448
See also Chernoff bounds; Hoeffding’s inequality
biological hot spot, 260
bipartite feedback, 343, 354
Bishop, Christopher M., 49
bits, 231
Blackwell, David, 169
Blum, Avrim, 297
Blumer, Anselm, 49
Boolean functions, approximability of, 73x
boost-by-majority
and AdaBoost, 448–452
algorithm, 427–428
analysis, 428–430
continuous-time limit, 462–468
generalization error of, 432–433
making adaptive, 460–462
with margin target, 454–455x
NonAdaBoost as limit of, 449–451
non-adaptiveness, 459
optimality of, 430–432
with randomized weak learner, 431, 458x
weighting function non-monotonic, 451
boosting, 4–5
and bias-variance analysis, 118–120
Cameron, A., 263
cancer genes, 366
found using RankBoost, 366–367
CART, 115
See also decision trees
Cartesian product, 491
cascading of classifiers, 70
Catlett, Jason, 133, 220
Čencov, N. N., 262
Censor, Yair, 262
census dataset, 129
and estimating conditional probabilities, 202
central limit theorem, 498–499
Cesa-Bianchi, Nicolò, 170
C4.5, 11, 14–15, 118, 290
See also decision trees
C4.5rules, 290
chain rule, 496
Chapelle, Olivier, 369
Cheamanunkul, Sunsern, 486
Chentsov, N. N., 262
Chernoff bounds, 30
and AdaBoost error, 448
optimality of, 448
See also Hoeffding’s inequality
Chervonenkis, A. Ya., 49
chip, 419
chip game, 419–420
approximating optimal play, 422–427
and online prediction, 455–456x
optimal play in, 420–422, 430–432
potential function, 426
relaxed versions, 431–432, 457–458x
clamping, 384–385
limited effect of, 393
classification exponential loss, 349
See also exponential loss
classification loss, 177
classification problem, learning, 2–4
classification rule, 3
classifier, 3
closed (set), 495
closure (of set), 495
code length, 231
codes, output. See output codes
codeword, 321
coding matrix, 321
Cohen, William W., 297
Cohn, David, 133–134
coin flipping, 30–31, 39–40
See also binomial distribution
Collins, Michael, 219, 262, 297, 369
combined classifier, 7
form of AdaBoost’s, 77
randomized, 72x
in support-vector machines versus AdaBoost, 126–127
compact (set), 495
complementary error function, 465
approximation of, 481
complexity. See simplicity
complexity measures
for finite hypothesis spaces, 34
VC-dimension, 37–38
compression achieved by AdaBoost, 88
compression-based analysis of AdaBoost, 83–86
compression schemes, 41–42
hybrid, 84–85
and VC-dimension, 91x
computation node (of feedforward network), 89x
conditional likelihood, 195
conditional probability, 194
as function, 378
conditional probability, estimating
caveats, 202
and convex program formulation, 253–255
with exponential loss, 202
with general loss functions, 200–201
with logistic regression, 194–195
and overfitting, 202–203
conditions sufficient for learning, 24–28
confidence
in active learning, 129–132
applications of, 128–132
measured by margin, 95
rejection with low, 128–129
confidence-rated AdaBoost, 273–274
for alternating decision trees, 293–294
convergence of, 300–301x
for rule-set learning, 287–289
training error of, 274
See also confidence-rated predictions
confidence-rated predictions
abstaining, 278–279 (see also abstaining weak hypotheses)
binary, 277
bounded range, 277–278
with domain partitioning, 283–285
dropping αt, 281–283
experiments, 286
general methods, 275–277
interpretation, 273
and margins analysis, 297–298x
motivating example, 271–272
smoothed, 284–285
conservation biology, 255–256
conservative learner, 455x
consistency (with training set), 24
improved error bounds, 39–40
consistency, statistical. See universal consistency
(statistical)
consistency, universal. See universal consistency (in online prediction); universal consistency
(statistical)
constrained optimization
and regularization, 204
See also convex program; linear programming
constraint selection
in AdaBoost, 235
cyclic, 229, 266x
greedy, 229
context trees, 166–167
continuous functions, 494–495
continuous-time limit of boost-by-majority, 462–468
differential equations for, 465–467
potential function, 463–467
weighting function, 467–468
contour map, 220x
convergence
of AdaBoost, 243–251
of AdaBoost.L, 265–266x
of AdaBoost to Bayes optimal, 386–387
almost, sure, 386
of confidence-rated AdaBoost, 300–301x
of coordinate descent, 182, 263x
of distributions, 251
with probability one, 386
rate for AdaBoost, 387–393
rate for AdaBoost.S, 408–410x
rate for AdaBoost with bounded updates, 410–411x
of unnormalized weight vectors, 248
convex duality, 251–252
convex hull, 97
Rademacher complexity of, 108–109
convexity, 496–497
convex loss
and Bayes error, 407–408x
poor accuracy if minimized, 412–413x
See also exponential loss; logistic loss
convex program, 228
for AdaBoost, 234, 239, 253–255
for density estimation, 258–259
for logistic regression, 253–255
coordinate descent, 179–184
and AdaBoost, 180–181
convergence of, 182, 263x
and functional gradient descent, 191–192
for general loss functions, 182–184
and gradient descent, 181–182
on logistic loss, 197
on square loss, 183–184
Cortes, Corinna, 133, 369
Cover, Thomas M., 262, 453
Cristianini, Nello, 133
cross validation, 13, 289
Csiszár, Imre, 262
cumulative loss (in repeated game), 146
bounds on, 147–151
curse of dimensionality, 126
cutoff, early (in BrownBoost), 470–471
Subject and Author Index

<table>
<thead>
<tr>
<th>Term</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cyclic behavior</td>
<td>239</td>
</tr>
<tr>
<td>example</td>
<td>263–264x</td>
</tr>
<tr>
<td>cyclic constraint selection</td>
<td>229</td>
</tr>
<tr>
<td>convergence</td>
<td>266x</td>
</tr>
<tr>
<td>data compression</td>
<td>169</td>
</tr>
<tr>
<td>data-limited learning</td>
<td>129–132, 211–219</td>
</tr>
<tr>
<td>datasets</td>
<td></td>
</tr>
<tr>
<td>See AP headlines dataset; benchmark</td>
<td></td>
</tr>
<tr>
<td>datasets; census dataset; heart-disease dataset; letter dataset; spoken-dialogue task</td>
<td></td>
</tr>
<tr>
<td>Daubechies, Ingrid</td>
<td>133, 262</td>
</tr>
<tr>
<td>decision stumps</td>
<td>13</td>
</tr>
<tr>
<td>algorithm for</td>
<td>64–66x</td>
</tr>
<tr>
<td>bias and variance of</td>
<td>118–120</td>
</tr>
<tr>
<td>for binary features</td>
<td>64–65</td>
</tr>
<tr>
<td>confidence-rated</td>
<td>284</td>
</tr>
<tr>
<td>consistency of voted</td>
<td>137x</td>
</tr>
<tr>
<td>for continuous features</td>
<td>66</td>
</tr>
<tr>
<td>for discrete features</td>
<td>65</td>
</tr>
<tr>
<td>growth function of</td>
<td>52x</td>
</tr>
<tr>
<td>and learning hyper-rectangles</td>
<td>59</td>
</tr>
<tr>
<td>VC-dimension of</td>
<td>52x</td>
</tr>
<tr>
<td>See also</td>
<td></td>
</tr>
<tr>
<td>threshold rules</td>
<td></td>
</tr>
<tr>
<td>decision trees</td>
<td>14–15</td>
</tr>
<tr>
<td>alternating</td>
<td>298–300x</td>
</tr>
<tr>
<td>bias and variance of</td>
<td>118–120</td>
</tr>
<tr>
<td>boosting-style analysis</td>
<td>300x</td>
</tr>
<tr>
<td>as domain-partitioning weak hypotheses</td>
<td>283</td>
</tr>
<tr>
<td>optimal risk of</td>
<td>412x</td>
</tr>
<tr>
<td>in penny-matching</td>
<td>166–167</td>
</tr>
<tr>
<td>uncontrolled complexity of</td>
<td>115</td>
</tr>
<tr>
<td>See also</td>
<td></td>
</tr>
<tr>
<td>CART; C4.5</td>
<td></td>
</tr>
<tr>
<td>Della Pietra, Stephen</td>
<td>262</td>
</tr>
<tr>
<td>Della Pietra, Vincent</td>
<td>262</td>
</tr>
<tr>
<td>delta functions</td>
<td>121, 137x</td>
</tr>
<tr>
<td>Demiriz, Ayhan</td>
<td>170, 220</td>
</tr>
<tr>
<td>density estimation</td>
<td>256–258</td>
</tr>
<tr>
<td>convex program for</td>
<td>258–259</td>
</tr>
<tr>
<td>derivative</td>
<td>495–496</td>
</tr>
<tr>
<td>Devroye, Luc</td>
<td>49</td>
</tr>
<tr>
<td>dichotomies</td>
<td>34</td>
</tr>
<tr>
<td>in output code</td>
<td>321</td>
</tr>
<tr>
<td>realized by AdaBoost’s combined classifier</td>
<td>78–79, 81–82</td>
</tr>
<tr>
<td>Dietterich, Thomas G.</td>
<td>120, 133, 333, 406</td>
</tr>
<tr>
<td>difference (of sets)</td>
<td>491</td>
</tr>
<tr>
<td>differential equations</td>
<td>465–467</td>
</tr>
<tr>
<td>direct approach</td>
<td>24–28</td>
</tr>
<tr>
<td>direct bounds for AdaBoost</td>
<td></td>
</tr>
<tr>
<td>See form-based analysis</td>
<td></td>
</tr>
<tr>
<td>discriminative approach</td>
<td>28–29</td>
</tr>
<tr>
<td>distribution modeling</td>
<td></td>
</tr>
<tr>
<td>See species distribution</td>
<td></td>
</tr>
<tr>
<td>modeling</td>
<td></td>
</tr>
<tr>
<td>document retrieval</td>
<td>354, 358</td>
</tr>
<tr>
<td>domain</td>
<td>3</td>
</tr>
<tr>
<td>domain-partitioning weak hypotheses</td>
<td>283–285</td>
</tr>
<tr>
<td>abstaining</td>
<td>293–294</td>
</tr>
<tr>
<td>multiclass</td>
<td>314</td>
</tr>
<tr>
<td>smoothed predictions for</td>
<td>284–285</td>
</tr>
<tr>
<td>Doshi, Anup</td>
<td>170</td>
</tr>
<tr>
<td>drifting games</td>
<td>453, 486</td>
</tr>
<tr>
<td>dual (of game)</td>
<td>159</td>
</tr>
<tr>
<td>value of</td>
<td>171x</td>
</tr>
<tr>
<td>dual form (of linear program)</td>
<td>173x</td>
</tr>
<tr>
<td>duality, convex</td>
<td>251–252</td>
</tr>
<tr>
<td>dual norms</td>
<td>492–493</td>
</tr>
<tr>
<td>dual optimization problem</td>
<td>497</td>
</tr>
<tr>
<td>Duda, Richard O.</td>
<td>49</td>
</tr>
<tr>
<td>Dudík, Miroslav</td>
<td>263</td>
</tr>
<tr>
<td>Dudley, R. M.</td>
<td>89</td>
</tr>
<tr>
<td>Duffy, Nigel</td>
<td>220</td>
</tr>
<tr>
<td>dynamics of AdaBoost</td>
<td>239</td>
</tr>
<tr>
<td>example</td>
<td>263–264x</td>
</tr>
<tr>
<td>early stopping</td>
<td>207</td>
</tr>
<tr>
<td>ECOC</td>
<td>322–323</td>
</tr>
<tr>
<td>See also</td>
<td></td>
</tr>
<tr>
<td>output codes</td>
<td></td>
</tr>
<tr>
<td>ecological niche</td>
<td>256</td>
</tr>
<tr>
<td>edges</td>
<td>54</td>
</tr>
<tr>
<td>and margins</td>
<td>112–114, 116–117, 158–159</td>
</tr>
<tr>
<td>efficiency, optimal</td>
<td>433–447</td>
</tr>
<tr>
<td>Ehrenfeucht, Andrzej</td>
<td>49</td>
</tr>
<tr>
<td>Eibl, Günther</td>
<td>333</td>
</tr>
<tr>
<td>elections</td>
<td>97</td>
</tr>
<tr>
<td>Elith, Jane</td>
<td>263</td>
</tr>
<tr>
<td>email, junk</td>
<td>1</td>
</tr>
<tr>
<td>empirical error</td>
<td>See training error</td>
</tr>
<tr>
<td>empirical ranking loss</td>
<td>345</td>
</tr>
<tr>
<td>empirical risk</td>
<td>379</td>
</tr>
<tr>
<td>of AdaBoost.S</td>
<td>408–410x</td>
</tr>
<tr>
<td>of AdaBoost with bounded updates</td>
<td>410–411x</td>
</tr>
<tr>
<td>rate minimized by AdaBoost</td>
<td>387–393</td>
</tr>
<tr>
<td>and true risk</td>
<td>393–396</td>
</tr>
<tr>
<td>empirical weak learning assumption</td>
<td>See weak learning assumptions</td>
</tr>
<tr>
<td>empty set</td>
<td>491</td>
</tr>
<tr>
<td>entropy</td>
<td>231</td>
</tr>
<tr>
<td>base of</td>
<td>232</td>
</tr>
<tr>
<td>maximum</td>
<td>234, 258–260</td>
</tr>
<tr>
<td>environment (in game)</td>
<td>146</td>
</tr>
<tr>
<td>adversarial</td>
<td>151</td>
</tr>
<tr>
<td>environmental variables</td>
<td>256</td>
</tr>
<tr>
<td>ε-AdaBoost</td>
<td></td>
</tr>
<tr>
<td>ε-boosting</td>
<td></td>
</tr>
<tr>
<td>See α-Boost</td>
<td></td>
</tr>
<tr>
<td>equilibrium margin</td>
<td>132</td>
</tr>
<tr>
<td>erf. See</td>
<td></td>
</tr>
<tr>
<td>complementary error function</td>
<td></td>
</tr>
<tr>
<td>error. See</td>
<td></td>
</tr>
<tr>
<td>Bayes error; generalization error; test error; training error</td>
<td></td>
</tr>
<tr>
<td>error, weighted</td>
<td>6</td>
</tr>
<tr>
<td>error analysis of AdaBoost</td>
<td></td>
</tr>
<tr>
<td>for AdaBoost.MO</td>
<td>334–337x</td>
</tr>
<tr>
<td>basic assumptions</td>
<td>75–77</td>
</tr>
<tr>
<td>compression-based</td>
<td>83–86</td>
</tr>
<tr>
<td>form-based (see form-based analysis)</td>
<td></td>
</tr>
</tbody>
</table>
error analysis of AdaBoost (cont.)
margins-based (see margins analysis)
with Rademacher complexity, 106–111
error bounds, 30–43
absolute, 43–46
for compression schemes, 41–42
for consistent hypotheses, 39–40
for countable hypothesis spaces, 51x
for finite hypothesis spaces, 32–34
and hybrid compression schemes, 84–85
for infinite hypothesis spaces, 34–38
looseness of, 43
lower bound for boosting, 433–447
lower bound for multiclass boosting, 458x
with Rademacher complexity, 107–108
for single hypothesis, 30–32
using growth function, 35–36
using union bound, 33
using VC-dimension, 37
See also margins analysis
error-correcting output codes, 322–323
See also output codes
error function, complementary. See complementary error function
Euclidean geometry
in iterative projection algorithms, 230
in support-vector machines, 122
Euclidean norms. See norms
example, 3
example weights, 62–63
exception list, 436
exhaustive AdaBoost, 185
exhaustive weak learning algorithm, 58
experiments
active learning, 130–132
on benchmarks, 11
with confidence-rated predictions, 286
on heart-disease dataset, 11–13
incorporating prior knowledge, 213–215
multiclass boosting, 308–309
noisy, with AdaBoost, 404
noisy, with BrownBoost, 484–485
with output codes, 332–333
penny-matching, 167–169
species distribution modeling, 260
exponential fictitious play, 169
exponential loss, 178
and AdaBoost, 177–179
in confidence-rated AdaBoost, 275
convex program for minimizing, 253–255
and functional gradient descent, 190–191
and generalization error, 184–188
and gradient descent, 185–186
and iterative projection algorithms, 244–246
versus logistic loss, 196
no finite minimum, 182
non-unique minimum, 186
poor accuracy if minimized, 398–404
provably minimized, 248
versus ranking exponential loss, 348–351
rate minimized by AdaBoost, 387–393
for semi-supervised learning, 216–217
See also risk
exponential weights. See MW
expression levels, 366
face detection, 66–70
and active learning, 130
and cascade of classifiers, 70
rectangular patterns for, 68
feasibility (of linear program), 173x
feasible set, 228
for AdaBoost, 233, 241
with inequality constraints, 266–267x
nonemptiness conditions, 237–239
features, 64, 194, 256–258
feature space, 194
feedback (for ranking), 343–344
bipartite, 343, 354
inconsistent, 343–344
layered, 343, 353–354
quasi-bipartite, 358
quasi-layered, 357–358
weighted, 357
feedback function, 357
feedback graph, 343
feedforward network, 89–91x
Fibonacci sequence, 263x
fictitious play, exponential, 169
final classifier. See combined classifier
Floyd, Sally, 49, 89
form-based analysis
for finite hypothesis spaces, 78–81
for infinite hypothesis spaces, 81–83
and overfitting, 80–81
Foster, Dean P., 169
Freund, Yoav, 71, 89, 169–170, 220, 297, 333, 369, 453, 486
Friedman, Jerome H., 170, 219–220, 333
Fudenberg, Drew, 169–170
functional, 188
functional gradient descent, 188–193
and AdaBoost, 190–191
with classification learner, 192–193
and coordinate descent, 191–192
on logistic loss, 197–200
with regression learner, 193
for semi-supervised learning, 217–218
on square loss, 193
Fürnkranz, Johannes, 297
Gale, William A., 220
game, voting. See voting game
games, learning in repeated, 145–151
and approximately solving a game, 152–153
and boosting, 159–163
and online prediction, 155–157
and proof of minimax theorem, 151–152
versus voting game, 417–418
See also MW games, matrix
bounded range, 145–146
defined, 142
minmax and maxmin strategies, 143–144
randomized play in, 142–143
sequential play in, 143–144
solving, 145
value of, 144–145
See also games, learning in repeated
games, repeated. See games, learning in repeated
games, zero-sum, 142
game theory, 142–145
and boosting, 157–163
game value, 144–145
of dual, 171x
and MW analysis, 151
Gaussian distribution, 28–29, 498
Gauss-Southwell, 191
Gautschi, Walter, 486
gene expression levels, 366
generalization error, 3, 26
absolute guarantees, 43–46
of AdaBoost (see error analysis of AdaBoost)
of boost-by-majority, 432–433
bounds on (see error bounds)
form-based analysis of (see form-based analysis)
and loss minimization, 184–188
margin-based analysis of (see margins analysis)
of support-vector machines, 91x
generalized AdaBoost. See confidence-rated AdaBoost
generalized output codes, 325–327
generative approach, 28–29
genes, cancer, 366
found using RankBoost, 366–367
Gentle, Claudio, 49
GentleAdaBoost, 223x
geometry. See Euclidean geometry; information geometry
Gorin, A. L., 333
gradient, 182, 495–496
gradient descent, 185, 221–222x
and coordinate descent, 181–182
on exponential loss, 185–186
See also functional gradient descent
greedy constraint selection, 229
ground hypothesis, 436
Grove, Adam J., 133, 170
growth function, 35–36
in abstract formulation, 38
of feedforward network, 90x
Günwald, Peter D., 49
Gubin, L. G., 262
Guruswami, Venkatesan, 333
Guyon, Isabelle M., 133
Györö, László, 49
Hadamard matrix, 222x
Hagelbarger, D. W., 163, 170
Hakkani-Tür, Dilek, 134
Halperin, I., 262
Hamming decoding, 322, 328
Hamming loss, 311–312
and one-error, 315
Hannan, James, 169
Hannan consistency, 169
“hard” distributions, 73x
hard predictions, 271–272
Hart, Peter E., 49
Hart, Sergiu, 169
Hastie, Trevor, 170, 219–220, 333
Haussler, David, 89
heart-disease dataset, 3–4
and alternating decision trees, 294–296
experiments with AdaBoost, 11–13
Helmold, David P., 49, 170, 220
hierarchy of classes, 327
Hoeffding, Wassily, 49, 71, 453
Hoeffding’s inequality, 30–31
and AdaBoost’s training error, 60–62
generalized, 438
proof of, 135–136x
See also Chernoff bounds
Höpfgen, Klaus-U., 220
Holte, Robert C., 71
hot spot, biological, 260
hybrid compression schemes, 84–85
applied to AdaBoost, 85–86
hyper-rectangles. See axis-aligned rectangles
hypothesis, 3
hypothesis class. See hypothesis space
hypothesis space, 32
complexity of, 34
cvx hull of, 97
span of, 382–383
if-then rules. See rule (if-then)
indicator function, 26, 491
inequality constraints, 266x
infimum, 493
information geometry, 234
information retrieval, 354, 358
information-theoretic measures, 230–232
See also binary relative entropy; entropy; relative entropy; unnormalized relative entropy
initial distribution, 54
inner product, 492
and kernels, 125
input node (of feedforward network), 89x
instance-based weak learner (for ranking), 352–353
instances, 3
instance space, 3
integral image, 69–70
intermediate value theorem, 495
intersection (of sets), 491
intervals, unions of, 36
learning, 51x
VC-dimension of, 36, 51x
iterative projection algorithms, 228–230
and AdaBoost, 232–237, 239–242
constraint selection, 229–230
with Euclidean geometry, 228–230
examples, 229
and exponential loss minimization, 244–246
geometry of solution, 243–244, 247–248
with inequality constraints, 266–268x
proof of convergence, 246–251
Jaynes, E. T., 262
Jell-O, 431–432
Jensen’s inequality, 497
Jevtić, Nikola, 333
Jiang, Wenxin, 406
Jones, Michael, 71
Jordan, Michael I., 406
Kalai, Adam Tauman, 406
Kapur, J. N., 262
Kearns, Michael J., 49–50, 297
kernels, 125, 128
Kesavan, H. K., 262
Kivinen, Jyrki, 262
Klautau, Aldebaro, 333
KL divergence. See relative entropy
Kohavi, Ron, 133, 297
Kolchinskii, V., 133
Kong, Eun Bae, 120, 133
Koo, Terry, 297, 369
k-partite feedback. See layered feedback
Kremen, C., 263
Kullback, Solomon, 262
Kullback-Leibler divergence. See relative entropy
Kunz, Clayton, 297
label, 3
Ladner, Richard, 133
Lafferty, John D., 262–263
Lagrange multipliers, method of, 497
Lagrangian, 497
Lane, Terran, 133
Langford, John, 333
large-margin instance, 398, 402
lasso, 220
See also regularization
layered feedback, 343, 353–354
lazy booster, 418
learner (in game), 146
learning, conditions for, 24–28
learning algorithm, 3
See also individual algorithms by name
learning rate, 189
learning repeated games. See games, learning in repeated
learning to rank. See ranking
least-squares regression. See linear regression
Lebanon, Guy, 263
Leibler, R. A., 262
letter dataset, 15
Levine, David K., 169–170
Lewis, David D., 133, 220
Li, Hang, 369
likelihood, 259
conditional, 195
lim inf, 494
limit, 493–494
lim sup, 494
linear constraints, 228
linear programming, 170
and games, 173–174x
linear regression, 175, 183
linear separability, 57–60
definition, 57–58
and online prediction, 172x
and weak learnability, 58–60, 116, 158–159
linear threshold functions, 38
and AdaBoost’s combined classifiers, 77
in support-vector machines, 122
in support-vector machines versus AdaBoost, 126–127
VC-dimension of, 77–78
See also voting classifiers
line search, 185
lingering (run of BrownBoost), 477
Lipschitz functions, 109, 136x
Littlestone, Nick, 49, 89, 169
Liu, Tie-Yan, 369
LogAdaBoost, 199n
logarithm, 492
logistic loss, 195
convex program for minimizing, 253–255
versus exponential loss, 196
modifying AdaBoost for, 197–200
logistic regression, 194–196
as convex program, 253–255
loss function for, 195
modifying AdaBoost for, 197–200
multiclass, 372–373x
unified with AdaBoost, 252–255
See also AdaBoost.L; logistic loss
LogitBoost, 199n, 223–224x
See also AdaBoost.L
log likelihood, 259
ℓ₁-norm, 204
Long, Philip M., 406
loss (in matrix games), 142
See also cumulative loss (in repeated game)
loss-based decoding, 322, 328
for all-pairs reduction, 337–338x
loss function, 175
of AdaBoost, 177–179
comparison of, 196
coordinate descent for general, 182–184
incorporating prior knowledge, 212–213
for semi-supervised learning, 216–217
See also Hamming loss; classification loss;
exponential loss; logistic loss; ranking loss; risk;
margin analysis, 97–106
and bagging, 120
and confidence-rated predictions, 297–298x
for finite hypothesis spaces, 98–104
interpretation of bounds, 98–99
minimum margin versus entire distribution, 115–116
multiclass, 334–337x
and overfitting, 99
using minimum margin, 106
using Rademacher complexity, 109–111
margins theory
and active learning, 129–132
applications of, 128–132
versus loss minimization, 187
and universal consistency, 397–398
See also margin maximization; margins; margins
analysis
martingales, 61
Marx, Groucho, 364
Mas-Colell, Andreu, 169
Mason, Llew, 133, 219, 297
matrix games. See games, matrix
maxent. See maximum entropy
maximum, 493
maximum entropy, 234
for species distribution modeling, 258–260
for support-vector machines versus AdaBoost, 127
for species distribution modeling, 260
maxmin strategy, 144
approximate, 153
and boosting, 161
McAllester, David, 406
McAuliffe, Jon D., 406
Mease, David, 219
medical diagnosis, 3–4
Meir, Ron, 406
microarray, 366
mind-reading game. See penny-matching
mind-reading machine, 165
“mini” boosting algorithm, 71–72x
minimizing (versus minimize), 6
minimum, 493
minimum description length principle, 49
minmax strategy, 143
approximate, 152
and boosting, 161
problems with playing, 145
proof of, 151–152, 171x
with pure strategies, 170x
mislabeled data. See noise
misorderings, 345
mistake-bounded learning. See online
prediction
Subject and Author Index

mistake matrix, 155
dual of, 159–160
value of, 157–159
mixed strategies, 143
model, 3
Mohri, Mehryar, 369
monomials, Boolean, 50x
Moon, Taesup, 369
movie ranking, 342
Mukherjee, Indraneel, 333, 406, 453
multiclass boosting
   based on ranking, 361
   experiments, 308–309
   lower bound on error, 458x
   and margins analysis, 334–337x
   weak learning assumptions, 305–308
See also AdaBoost.MH; AdaBoost.MO;
   AdaBoost.MR; AdaBoost.M1; multiclass-to-binary
   reductions
   multiclass-to-binary reductions, 303–304
   all-pairs, 325
   with generalized output codes, 325–328
   with hierarchy of classes, 327
   one-against-all, 303, 311–313
   with output codes, 320–322
   multi-label classification, 310
   multilayer perceptron, 91x
   multiplicative weights. See MW
MW, 146–147
   and α-Boost, 162
   analysis of, 147–151
   and approximately solving a game, 152–153
   for boosting, 159–162
   and game value, 151
   for online prediction, 155–156
   and proof of minmax theorem, 151–152
   self-play with, 171x
   setting parameters of, 149–150
   with varying parameter values, 171–172x
natural languages, ambiguity of, 364
neural network, 91x, 132
Newton’s method, 223x
niche, 256
noise
   effect on AdaBoost, 398, 404–405
   experiments with AdaBoost, 404
   experiments with BrownBoost, 484–485
   handling, 405
NonAdaBoost, 419
   continuous-time limit of, 487x
   as limit of boost-by-majority, 449–451
   non-adaptive boosting. See NonAdaBoost
   normal distribution, 28–29, 498
   normal form, 142
   normalization factor, 179
   in confidence-rated AdaBoost, 275
norms, 492–493
   of functions in span, 383–384
   in support-vector machines versus AdaBoost, 127
notation, general, 491–492
NP-completeness, 177
numerical difficulties, 280n
O (big Oh notation), 492
objective function. See loss function
oblivious weak learner, 418
   optimality of, 430–432
   and potential function, 430–431
   Occam’s razor, 13–14
   odds and evens. See penny-matching
   one-against-all reduction, 303, 311–313
   one-error, 310–311
   of AdaBoost.MH, 314–315
   of AdaBoost.MR, 361–363
   and Hamming loss, 315
online learning. See online prediction
   online prediction, 153–157
   versus batch, 153–154
   as chip game, 455–456x
   as game playing, 155–157
   and linear separability, 172x
   model for, 154–155
   and penny-matching, 165–167
   and Rademacher complexity, 171x
Opitz, David, 406
Opper, Manfred, 486
   optimal boosting
   lower bound, 433–447
   See also boost-by-majority
   optimal edge, 116
   optimal encoding, 231
   optimal margin, 116
   optimal play (in chip game), 420–422,
   430–432
   approximated, 422–427
   optimal predictor, 379–380
   and Bayes optimal classifier, 380
   optimal risk, 379–380
   approached by AdaBoost, 385–387
   approached by decision trees, 412x
   approached by functions in span, 383
   and Bayes error, 380–382
   and regularization, 411–412x
   optimal strategy. See maximin strategy; minmax
   strategy
   optimization problem. See convex program; linear
   programming
   option trees, 297
   ordered pair, 492
   Orlitsky, Alon, 333
outliers
   and boost-by-majority, 451
detection, 317
output codes, 320–322
- based on hierarchy, 327
design of, 322–323, 326–327
- error-correcting, 322–323
- experiments with, 332–333
generalized (ternary), 325–327
output node (of feedforward network), 89x
overfitting, 13–14
- AdaBoost’s resistance to, 15–16
- and estimating conditional probabilities, 202–203
- and margins analysis, 99
- and theoretical bounds, 42
- and universal consistency, 397
of voting classifiers, 120–122, 137x
Oza, Nikunj C., 170
PAC learning, 44–47
- and computational intractability, 45–46
- equivalence of strong and weak, 46–47, 86–88
- general resource requirements, 88
- strong, 45
- weak, 46
pair, ordered, 492
pair-based weak learner (for ranking), 353
Panchenko, D., 133
Parsimonious play, 141, 143
Pereira, Fernando, 333
Phillips, Steven J., 263
polling, 97, 99
Polyak, B. T., 262
position (of chip), 419
in continuous time, 464
potential function
- of BrownBoost, 465
for chip game, 426
in continuous-time limit, 463–467
- discontinuity of BrownBoost’s, 467, 486x
in MW analysis, 147
and random play of oblivious weak learner, 430–431
power set, 491
prediction node, 291
prediction rule, 3
predictor, 3
preference pairs, 343
presence-only data, 256
primal form (of linear program), 173x
prior knowledge, incorporating, 211–215
prior model, 212
probabilistic method, 435
probability density function, 498
probably approximately correct. See
- PAC learning
projection, 228
- projection, iterative. See iterative projection
algorithms
Pujol, Oriol, 333
puller instance, 398, 402
pure strategies, 143
- and minmax theorem, 170x
Pythagorean theorem, 246–247
quadratic loss. See square loss
quasi-bipartite feedback, 358
quasi-layered feedback, 357–358
Quinlan, J. Ross, 297
Rademacher complexity, 106–111
- alternative definition, 107n
- for classifiers, 108, 171x
- and error bounds, 107–108
- and Lipschitz functions, 109, 136x
- and margins analysis, 109–111
- and support-vector machines, 136–137x
of voting classifiers, 108–109
Radeva, Petia, 333
Raik, E. V., 262
Rajaram, Shyamsundar, 369
random AdaBoost, 185–186
random forests, 133
randomized play, 142–143
randomized predictions, 72x
random projections, 137–138x
random variables, unbounded, 384
RankBoost, 345–348
and AdaBoost.MO, 369–370x
- based on reduction, 346
- with binary weak learner, 351–353
- choosing αt in, 347–348
- and confidence-rated AdaBoost, 370x
- criterion to optimize, 347–348
- for finding cancer genes, 366–367
- for graded feedback, 370–371x
- for layered feedback, 354–355
- loss compared to AdaBoost, 348–351
- for multiclass classification, 361
- for multiclass logistic regression, 372–373x
- for parsing, 365–366
- for quasi-layered feedback, 358–359
- ranking loss of, 347
- See also weak learner (for ranking)
RankBoost.L, 354–355
RankBoost.LQ, 358–359
- for multiclass classification, 361
ranked retrieval, 354, 358
ranking
- boosting for (see RankBoost)
feedback, 343–344
 ranking (cont.)
  framework, 342–345
  inconsistent feedback, 343–344
  for multiclass classification, 361
  for multiclass logistic regression, 372–373x
  reduction to binary, 346
  ranking exponential loss, 349
  versus exponential loss, 348–351
  ranking loss, 344–345
  of RankBoost, 347
  Rätsch, Gunnar, 71, 133
  Ravikumar, Pradeep, 333
  real-valued weak hypotheses. See confidence-rated
  predictions
  receiver-operating-characteristic curve, 369
  rectangles, axis-aligned. See axis-aligned rectangles
  reductions
  boosting to repeated game, 160
  online prediction to repeated game, 155–156
  ranking to binary, 346
  See also multiclass-to-binary reductions
  reference distribution, 147
  reference function, 383–384
  regression. See linear regression; logistic regression
  regret, 169
  See also cumulative loss (in repeated game); online
  prediction
  regularization, 204–205
  and boosting, 205–209
  for density estimation, 259
  and margin maximization, 209–211
  properties of solutions, 224–225x
  and true risk, 411–412x
  regularization path, 205
  and trajectory of AdaBoost, 209
  and trajectory of α-Boost, 207–209
  rejecting low-confidence predictions, 128–129
  relative entropy, 147, 231–232
  base of, 232
  in MW analysis, 147–149
  See also binary relative entropy; unnormalized
  relative entropy
  relative loss. See cumulative loss (in repeated game);
  online prediction
  repeated games. See games, learning in repeated
  resampling, 62–63, 62n
  reserve design, 260
  residual, 183
  reverse index, 280
  reweighting, 63
  Reyzin, Lev, 133
  Richardson, Thomas, 220
  Ridgeway, Greg, 220
  RIPPER, 290
  risk, 201
  empirical, 379
  for general loss, 407–408x, 412–413x
  optimal (see optimal risk)
  optimal predictor, 379–380
  poor accuracy if minimized, 398–404, 412–413x
  true, 379
  See also exponential loss; loss function
  Ritov, Ya’acov, 406
  ROC curve, 369
  Rockafellar, R. Tyrrell, 262
  Rock-Paper-Scissors, 142
  and minmax theorem, 145
  and The Simpsons, 145
  Rosset, Saharon, 170, 220
  Rudin, Cynthia, 133, 262–263, 369, 406
  rule (if-then), 287
  condition of, 288–289
  examples of, 289
  rule (prediction), 3
  rule of thumb, 1–2
  rule-sets, 287–290
  boosting algorithm for, 287–289
  other algorithms for, 297
  Russell, Stuart, 170
  Sahai, Amit, 333
  Sauer, N., 49
  Sauer’s lemma, 37
  tightness of, 51x
  Schapire, Robert E., 71, 89, 132–134, 169–170,
  Schohn, Greg, 134
  Schölkopf, Bernhard, 133
  Schuurmans, Dale, 133, 170
  seed, 436
  semantic classification, 316–317
  rejecting low-confidence predictions, 128
  semi-supervised learning, 215–219
  Sengupta, Shiladitya, 369
  sequence extrapolating robot, 163
  sequences, convergence of, 493–494
  sequential play, 143–144
  Servedio, Rocco A., 406
  set difference, 491
  Shalev-Shwartz, Shai, 71, 133
  Shannon, Claude E., 163, 170, 262
  shattering, 36
  Shawe-Taylor, John, 133, 170
  Shields, Paul C., 262
  Shhtarkov, Yuri M., 170
  sigmoid function, 194
  sign function, 491
  avoiding redefinition, 89x
  Simon, Hans-U., 220
  simplicity, 24–26
  Simpsons, The, 145
  Singer, Yoram, 71, 133, 219, 262, 297,
  333, 369
  Slater, 289–290
  Smola, Alex, 133
  smoothed predictions, 284–285
smooth margin, 134–135x
soft Oh, 492
solution, BrownBoost, 471–475, 488x
solving a game, 145
approximately, 152–153
and linear programming, 173–174x
spam filtering, 1
rejecting low-confidence predictions, 128
span, 382–383
sparsity
in abstaining weak hypotheses, 279–281
of approximate minmax strategy, 153
of output codes, 328
specialist, 297
species distribution modeling, 255–260
convex program for, 258–259
as density estimation, 256
experiments, 260
splitter node, 291
spoken-dialogue task, 316–317
and active learning, 130–132
limited data, 211
rejecting low-confidence predictions, 128
square loss, 175
coordinate descent on, 183–184
stability. See bias-variance analysis
standard deviation, 498
standard normal, 498
state of play, 165
“statistical view” of boosting, 219–220
stochastic coordinate descent. See random AdaBoost
stochastic differential equations, 486
Stork, David G., 49
strong learnability, 45
equivalent to weak learnability, 86–88
subsequence, convergent, 495
support-vector machines, 122–128
and boosting, 126–128
generalization error of, 91x
kernel trick, 125
with linearly inseparable data, 123
mapping to high dimensions, 123–125
and margin maximization, 122–123
and Rademacher complexity, 136–137x
VC-dimension of, 126
supremum, 493
SVMs. See support-vector machines
symmetric difference, 311
tail bounds. See binomial distribution
target class, 44
target error, 476
target function, 44, 154
Taylor’s theorem, 495
and empirical risk, 410–411x
term, 316
ternary output codes, 325–327
test error, 3
test examples, 3
test set, 3
Thomas, Joy A., 262, 453
three-round boosting algorithm, 71–72x
threshold rules, 24
compression scheme for, 41
finding best, 27
labelings induced by, 34
VC-dimension of, 36
See also decision stumps
Tibshirani, Robert, 219–220, 333
time, continuous. See continuous-time limit of
boost-by-majority
Tjalkens, Tjalling J., 170
top-down decision-tree algorithm, 298–300x
training error, 3, 26
of AdaBoost, 54–56
of AdaBoost.MH, 315–316, 324
of AdaBoost.MO, 323–325, 328–332
of AdaBoost.M1, 306–307
of AdaBoost.MR, 363
as biased estimate of generalization error, 26–27
of boost-by-majority, 428–430
of BrownBoost, 475–476
of confidence-rated AdaBoost, 274
looseness of bounds, 56
and randomized predictions, 72x
training examples, 3
training instances (for ranking), 342
as tuple, 24
trajectory of boosting, 205–207
and regularization path, 207–209
transpose (of matrix), 492
Traskin, Mikhail, 406
ture error. See generalization error
ture risk, 379
and empirical risk, 393–396
optimal (see optimal risk)
and regularization, 411–412x
tuples, 491
Tur, Gokhan, 134
unbounded random variables, 384
uniform convergence bounds, 32–33
abstract formulation, 38–39
for exponential loss (risk), 393–396
See also error bounds
union (of sets), 491
union bound, 31
and generalization error bounds, 33
unions of intervals. See intervals, unions of
universal consistency (in online prediction), 169
universal consistency (statistical), 378
of AdaBoost, 386–387
counterexample with binary predictions, 401–404
counterexample with confidence-rated predictions, 398–401
and margins theory, 397–398
and overfitting, 397
unlabeled data
in active learning, 129–132
in semi-supervised learning, 215–219
unnormalized relative entropy, 239
upper limit, 494
Valiant, Leslie G., 49–50
Vandenberghe, Lieven, 262
Vapnik, Vladimir N., 49, 133
Vapnik-Chervonenkis dimension. See VC-dimension
Vayatis, Nicolas, 406
Vazirani, Umesh V., 49
VC-dimension, 36
of affine threshold functions, 89x
of axis-aligned rectangles, 50x
as complexity measure, 37–38
and compression schemes, 91x
of decision stumps, 52x
of feedforward network, 90x
of finite hypothesis space, 50x
and generalization error, 37
of linear threshold functions, 77–78
and Rademacher complexity, 108
of support-vector machines, 126
Viola, Paul, 71
vMW, 171–172x
Vohra, Rakesh, 169
von Neumann, John, 169, 262
von Neumann minmax theorem. See minmax theorem
voting classifiers
in boost-by-majority, 461
and elections, 97
more complex than constituents, 120–122
Rademacher complexity of, 108–109
See also linear threshold functions
voting game, 417
versus repeated game, 417–418
Vovk, Volodimir G., 169–170
Warmuth, Manfred K., 49, 71, 89, 133, 169, 262
weak classifier. See weak hypotheses
weak hypotheses, 4
abstaining (see abstaining weak hypotheses)
complexity of, 76
domain-partitioning, 283–285
real-valued (see confidence-rated predictions)
sparse, 279–281
weak learnability, 46
and empty feasible set, 237–239
equivalent to strong learnability, 86–88
and linear separability, 58–60, 116, 158–159
and minmax theorem, 157–159
sufficient conditions for, 56–60
See also weak learning assumptions
weak learner (for classification). See weak learning algorithms
weak learner (for ranking)
design of, 347–348
instance-based, 352–353
pair-based, 353
weak learning algorithms, 4, 62–70
for decision stumps, 64–66
design approaches, 63–64
example weights used by, 62–63
exhaustive, 58
oblivious, 418
using random projections, 137–138x
weak learning assumptions, 4, 47–48
effect on boosting, 112–113
empirical, 48
and generalization error, 80–81
for multiclass, 305–308
in PAC model, 46
and training error of AdaBoost, 56
See also weak learnability
weak learning condition. See weak learning assumptions
weighted error, 6
weighted feedback, 357
Weighted Majority Algorithm, 154
weighted majority–vote classifiers. See voting classifiers
weighting function, 427
of BrownBoost, 467–468
in continuous-time limit, 467–468
weights. See \( a_t \), choosing; example weights
Widmer, Gerhard, 297
Wilcoxon–Mann–Whitney statistic, 369
Willems, Frans M. J., 170
Wyner, Abraham J., 219–220
Xi, Yongxin Taylor, 170
Xu, Jun, 369
Ye, Yinyu, 220
Yu, Bin, 170, 220, 262, 406
Zadrozný, Bianca, 333
Zakai, Alon, 406
Zenios, Stavros A., 262
0-1 loss. See classification loss
zero-sum games, 142
Zhang, Tong, 170, 262, 406
Zhang, Zhiheng, 219
Zhao, Peng, 220
Zhu, Ji, 170, 220, 333
Zonation, 260