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
α-factor, 479, 512
α-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
χ^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
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
χ^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
α-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
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
ρ-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
Θ_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