

Index

Note: Page numbers followed by *f* indicate figures and *t* indicate tables.

A

- Active constraints, 544
- Active set, 692
- Adaptive algorithm, 162
- Adaptive Boosting (AdaBoost) algorithm, 307–311
- Adaptive coordinate descent scheme, 259–261
- Adaptive CoSaMP (AdCoSaMP) algorithm, 479–480, 484*f*
- Adaptive decision feedback equalization, 202–204
- Adaptive gradient (ADAGRAD) algorithm, 368
- Adaptive line element (ADALINE), 881
- Adaptive projected subgradient method (APSM), 349–350
 - algorithm, 350
 - asymptotic consensus, 358
 - combine-then-adapt diffusion, 357–358
 - constrained learning, 356
 - convergence of, 351–356
 - distributed algorithms, 357–358
 - hyperslabs, 352
 - parameters, 352–356
 - projection operation, 351
 - SpAPSM, 480–484, 481*f*, 484*f*
- Adaptive signal processing, 5
- Adapt-then-combine DiLMS, 215–216, 216*f*
- Additive models approach, 568–570
- Ad hoc networks, 210
- ADMM algorithm. *See* Alternating direction method of multipliers (ADMM) algorithm
- Affine projection algorithm (APA), 188–194, 201
 - convergence, 353
 - curves for, 355*f*
 - geometric interpretation of, 189–191
 - normalized LMS, 193–194
 - orthogonal projections, 191–194
 - set-membership, 354
 - widely-linear, 195–196
- Affine set, 415–416
- Airborne Visible/Infrared Imaging Spectrometer (AVIRIS), 696–699
- ALMA algorithm, 560
- Alternating direction method of multipliers (ADMM)
 - algorithm, 220–221, 387–388
- Alternating optimization, 608–609
- Amino-acids, proteinogenic, 314–315
- Amplitude beam-pattern, 149*f*
- Analog signal
 - Fourier transform, 435*f*
 - sampling process, 432*f*, 434–435
- Analog-to-information sampling, 434–435
- Analysis of variance (ANOVA), 570
- APA. *See* Affine projection algorithm (APA)
- Approximate inference
 - block methods, 809–813
 - loopy belief propagation, 813–816
 - variational methods, 804–809
- Approximation error, 94, 374–376, 511
- APSM. *See* Adaptive projected subgradient method (APSM)
- Arithmetic averaging rule, 305
- ARMA model. *See* Autoregressive-moving average (ARMA) model
- AR models. *See* Autoregressive (AR) models
- Assumed density filtering (ADF), 682
- Asymptotic distribution, of LS estimator, 238–239
- Augmented Lagrangian method, 387–388
- Authorship identification, 570–573
- Autocorrelation matrix, 114–115
- Autocorrelation sequence, 33–35
- Auto-cumulants, 1021
- Autoencoders, 919–920, 925–927
- Automatic relevance determination (ARD), 655–656
- Autoregressive hidden Markov model, 829
- Autoregressive (AR) models, 38–40
- Autoregressive-moving average (ARMA) model, 40
- Autoregressive process estimation, 153
- Auxiliary particle filtering, 862–868
- Auxiliary variable Markov chain Monte Carlo methods, 735
- Average mutual information, 44–45, 47
- Average risk, Bayesian classification, 278–280
- Averaging method, 187
- Averaging rule, 217

B

- Backpropagation algorithm, 877, 886–897
 - activation functions, 894
 - cost function, 896–897
 - gradient descent scheme, 887–894
 - initialization, 894

- Backpropagation algorithm (*Continued*)
 - preprocess input variables, 894
 - target values, 894
- Backtracking, 783–785
- Backward errors, 138–140
- Backward MSE optimal predictors, 134–138
- Bagging, 303–304
- Bag-of-words approach, 571
- Bandpass filter, 37–38
- Base learner, 307
- Base transition matrices, 724
- Basis pursuit, 407, 418
- Basis pursuitde-noising (BPDN), 408
- Basis vector, 395
- Batch learning, 376–379
- Batch processing methods, 162
- Baum-Welch algorithm, 827–828
- Bayesian approach, 5
 - to regression, 589–593
 - to sparsity-aware learning, 655–661
- Bayesian classification, 276–280
 - average risk, 278–280
 - designing classifiers, 282
 - equiprobable Gaussian classes, 284f
 - Gaussian distributed classes, 283f
 - implicitly forms hypersurfaces, 281f
 - M -class problem, 278–279, 284, 293–294
 - misclassification error, 277–280
 - reject option, 279
- Bayesian decision theory, 276
- Bayesian inference, 84–89, 87f
- Bayesian information criterion (BIC), 599
- Bayesian learning, 11–12
 - neural networks, 902–903
 - regularization, 75
 - variational approximation, 640–645
- Bayesian networks (BNs)
 - causality, 753–755
 - cause-effect relationships, 753–755
 - completeness, 761–762
 - d -separation, 755–758, 758f
 - faithfulness, 761–762
 - graphs, 749–753
 - I-maps, 761–762
 - independent variables, 751f
 - joint pdf, 836–837
 - Kalman filtering, 852–854
 - latent Markov model, 818f
 - linear Gaussian models, 759–760
 - multiple-cause networks, 760, 761f
 - naive Bayes classifier, 835f
 - set of findings and diseases, 806f
 - sigmoidal, 758–759, 759f, 760
 - soundness, 761–762
 - triangulated graph, 800–801, 800f
- Bayesian regression, 690–692
 - computational considerations, 692
 - hyperparameters, 691–692
- Bayes's theorem, 13, 276–277, 586
- Beamforming, 145–148
- Belief propagation, 782
- Bernoulli distribution, 18
- Best linear unbiased estimator (BLUE), 144–145, 237–238
- Beta distribution, 25–26
- Bethe entropy approximation, 813–814
- Bethe free energy cost, 813–814
- Between-classes scatter matrix, 296
- Biased estimation, 64–67
- Biasor, 57–58
- Bias-variance dilemma/tradeoff, 77–81
- BIC. *See* Bayesian information criterion (BIC)
- Big data problems, 162
- Big data tasks, 376
- Binary classifier, 307–308
- Binomial deviance function, 311–313
- Binomial distribution, 18–19
- Bipartite graph, 769
- Blind source separation (BSS), 964–965
- Blocking Gibbs sampling, 734
- Block methods, 809–813
- Block processing techniques, 197
- Block sparsity, 468
- BLUE. *See* Best linear unbiased estimator (BLUE)
- BNs. *See* Bayesian networks (BNs)
- Boltzmann machines, 767
 - graph nodes representing, 812f
 - mean field approximation, 810–813
 - MRF, 809f
 - variational approximation, 807–809
- Boolean approach, bag-of-words, 571
- Boosting approach, 307–313
- Boosting trees, 313–314
- Bootstrap Aggregating, 303–304
- Bootstrap techniques, 93
- Box-Müller method, 713–714
- Bregman divergence, 389
- Burn-in phase, Metropolis method, 730

C

- Calculus of variations, 641
- Canonical correlation analysis (CCA)
 - content-based image retrieval, 953
 - correlation coefficient, 951–952
 - eigenvalue-eigenvector problem, 952, 953–954
 - goals, 951
 - optimization task, 951–952
 - PLS method, 954–955
- Capon beamforming, 148
- CART. *See* Classification and regression trees (CART)
- Cauchy-Schwartz inequality, 34, 396
- Cauchy sequence, 397
- Causality, 753–755
- Cause-effect relationships, 753–755
- Centralized networks, 209, 210^f
- Central limit theorem, 24–25
- Chain graph, 773–776
- Change-point detection, 737–738
- Channel equalization, 126–132
- Channel identification, 144–145
- Characteristic functions, 1020
- Chinese restaurant process (CRP), 683–684
- Chi-squared distribution, 27
- Cholesky factorization, 140, 255
- Circular condition, 115–116
- Class assignment rule, 303
- Classification, 2–3, 60–64
 - Bayesian classification, 276–280
 - decision (hyper)surfaces, 280–290
 - discrete nature, 275–276
 - Gaussian random process, 692–693
 - generative vs. discriminative learning, 63–64
 - logistic regression model for, 662–666
 - M*-class problem, 278–279, 284, 293–294
 - POCS, 347–349
 - protein folding prediction, 316, 318
 - trees, 300–304, 301^f
 - two-class task, 277, 286, 290^f, 296–297, 312
 - unstable, 303
- Classification and regression trees (CART), 300, 317^f, 318
- Classifiers, 2, 3^f
 - combining, 304–307
 - experimental comparisons, 304–305
 - goal to design, 60–61
 - schemes for combining, 305–307
- Class imbalance problem, 552
- Class label variable, 61
- Clifford algebras, 552
- Clique, 764^f, 769^f
 - message passing, 801–804
 - potentials, 763
- Closed convex set, 345–346
 - associated projection operator, 338–339
 - finite number, 341
 - Hilbert space, 334, 337, 339–340, 344
 - infinite, 349–356
 - nonempty intersection, 342
 - sliding window, 349–350
- Clustering, 3, 64, 617–620
- Cocktail party problem, 963–966
- Codon, 314–315
- Collapsed Gibbs sampling, 735
- Combine-then-adapt diffusion APSM, 357–358
- Combine-then-adapt diffusion LMS, 217
- Common sense reasoning, 753–754
- Communications channel, 43–44
- Compatibility functions/factors, 762
- Complementary slackness conditions, 1026
- Complete dictionaries, 414–415
- Complex linear space, 394
- Complex networks, 211
- Complex random variables, 16–17
- Complex-valued case
 - adaptive decision feedback equalization, 202–204
 - least mean fourth algorithm, 196–197
 - mean-square-error loss function, 175–176, 194–195
 - sign-error LMS, 196
 - transform-domain LMS, 197–201
 - widely-linear APA, 195–196
 - widely-linear LMS, 195
- Complex-valued data, widely linear RLS, 254–255
- Complex-valued variables
 - extension to, 111–118
 - widely linear, 113–116
 - Wirtinger calculus, 116–118
- Composite mirror descent, 389
- Compressed sensing (CS), 404
 - analog-to-information conversion, 434–436
 - definition, 430–431
 - description, 431–436
 - dimensionality reduction, 433–434
 - sparse signal representation, 487–488
 - stable embeddings, 433–434
 - sub-Nyquist sampling, 434–436
- Compressed sensing matching pursuit (CSMP) algorithms, 455–456, 460, 479–480

- Compressive sampling matching pursuit (CoSaMP), 455–456, 480
- Computational considerations, Bayesian regression, 692
- Computation, of lower bound, 650–651
- Concave, 667, 668–669
- Concentration parameter, 684
- Conditional entropy, 45
- Conditional independencies, 749, 752–753
- Conditional information, 43–44, 47
- Conditional log-likelihood, 835–836
- Conditional pdf, 632–633, 634–637
- Conditional probabilities, 12–13
- Conditional random fields (CRFs), 767–768
- Conditional Random Markov Field, 768
- Conditional restricted Boltzmann machine (CRBM), 920–922
- Conjugate function, 666–667
- Conjugate prior, 89
 - Dirichlet distribution, 604
 - gamma distribution, 601
 - Gaussian-gamma form, 603
- Conjugate Wirtinger’s derivative, 117
- Consensus-based algorithms, 221
- Consensus-based distributed schemes, 220–222
- Consensus matrix, 223–224
- Consensus strategy, 221–222
- Consistent estimator, 31
- Constrained-based path, 837
- Constrained learning, 356
- Constrained linear estimation, 145–148
- Continuous random variables, 14
 - average mutual information, 47
 - conditional information, 47
 - entropy for, 46–47
 - generalization, 45
 - Kullback-Leibler divergence, 47–48
 - relative entropy, 47–48
- Continuous-time signal, 29
- Continuous variables
 - beta distribution, 25–26
 - central limit theorem, 24–25
 - Dirichlet distribution, 27–49
 - exponential distribution, 25
 - gamma distribution, 26–27
 - Gaussian distribution, 20–24
 - uniform distribution, 20
- Contrastive divergence (CD), 911
- Convergence
 - affine projection algorithm, 353
 - APSM, 351–356
 - connection, 756
 - distributed learning, 181–186, 218–219
 - distributions, 49–51
 - error vector, 181–186
 - issues, Metropolis method, 731–732
 - in mean, 182–183, 218
 - NORMA, 559–560
 - performance, 218–219
 - stochastic (*see* Stochastic convergence)
- Convex, 330–333
 - duality, 666–671
 - online learning, 367–370
 - optimization techniques, 458, 460, 478, 487
 - programming, 1028–1029
 - separating hyperplane, 370
 - strictly, 331
 - theory, 328
- Convex set, 329
 - closed (*see* Closed convex set)
 - Hilbert space, 334, 337
 - strongly attracting nonexpansive mapping, 356
 - theory, 328
- Convolution matrices, 132
- Coordinate descent (CD), 258–261, 459
- Correlated component analysis, 953
- Correlation, 15
- Correlation matrix, 15–16
- CoSaMP. *See* Compressive sampling matching pursuit (CoSaMP)
- Cosparsity, 488–490
- Cost function
 - backpropagation algorithm, 896–897
 - isovalue curves, 164^f
 - surface, 107–108, 109^f
 - two-dimensional parameter space, 164^f
- Countably infinite, 683
- Coupon collector’s problem, 992
- Covariance, 15
 - functions, 688–689
 - Kalman algorithm, 152
 - matrix, 15–16, 175^f
- Cover’s theorem, 514–517
- Cramér-Rao bound, 67–72, 1019
- CRFs. *See* Conditional random fields (CRFs)
- Cross-correlation vector, 128, 129, 171–174
- Cross-cumulants, 1021
- Cross-entropy cost function, 896
- Cross-entropy error, 292
- Cross-spectral density, 120–121

Cross-validation, 92–93
 CS. *See* Compressed sensing (CS)
 CSMP algorithms. *See* Compressed sensing matching pursuit (CSMP) algorithms
 C-SVM, 547
 Cumulant generating function, 815
 Cumulants, 1020–1021
 Cumulative distribution function (cdf), 14, 19*f*, 713*f*
 Cumulative loss, 186–188, 371
 Cuprite data set, 696–697
 Curse of dimensionality, 89–91, 90*f*
 Curve fitting problem, 54–55, 55*f*
 Cyclic coordinate descent (CCD), 258–261
 Cyclic path, 210

D

DAG. *See* Directed acyclic graph (DAG)
 Dantzig selector, 472
 Data sets, 91
 De-blurring, 4–5, 4*f*
 Decentralized networks, 210–211
 Decision feedback equalizer (DFE), 202–203, 203*f*, 204*f*
 Decision surface, 60–61, 280–281, 282
 Gaussian distribution, 282–287
 naive Bayes classifier, 287–288
 nearest neighbor rule, 288–290
 Decision trees, 304
 CART, 317*f*
 protein folding prediction classification, 318
 Decomposition, analysis of variance, 570
 Deconvolution, 121–124, 126–132
 Deep belief network (DBN), 916–918, 928
 Deep learning, 877
 block diagram, 906*f*
 character recognition, 923–925
 CRBM, 920–922
 Gaussian visible units, 918–919
 issues, 903
 stacked autoencoder, 919–920
 training, 905–908
 Deflation procedure, 954–955
 Degeneracy phenomenon, 858–860
 Degree of node k , 211–212
 Deming regression, 262
 De-noising, 438–439, 439*f*
 Denoising autoencoder, 920
 Density function, 14
 Dependent random variable, 57–58

DFE. *See* Decision feedback equalizer (DFE)
 Dictionary learning (DL), 414–415
 codebook update, 967–968
 image de-noising, 970, 971*f*
 optimization problem, 966
 sparse coding, 967
 Difference equation, 38–39
 Diffusion gradient descent, 215
 Diffusion LMS (DiLMS), 211–218
 adapt-then-combine, 215–216, 216*f*
 combine-then-adapt, 217
 Dimensionality reduction, 243–244, 433–434
 Directed acyclic graph (DAG), 749
 Bayesian network, 749, 751
 d -separation, 758*f*
 independencies, 762*f*
 moralization on, 772*f*
 Directed graphs, 749, 772
 Dirichlet distribution, 27–49, 603–604
 Dirichlet process (DP), 684–686
 Discrete cosine transform (DCT), 412–413, 413*f*
 Discrete distributions
 cumulative distribution function, 713*f*
 generating samples from, 711–712
 resampling, 847–849
 Discrete random variables, 12–13
 codewords, 42–43
 entropy/average mutual information, 44–45
 information, 42–43
 mutual/conditional information, 43–44
 Discrete-time random process, 29
 Discrete-time stochastic process, 29*f*
 Discrete variables
 Bernoulli distribution, 18
 binomial distribution, 18–19
 multinomial distribution, 19–20
 Discrete wavelet transform (DWT), 412–413, 414–415
 Discriminant functions, 282
 Discriminative learning
 generative vs., 63–64
 hidden Markov model, 828–829
 Disjoint subsets, 302–303
 Distributed learning
 consensus-based schemes, 220–222
 convergence, 181–186, 218–219
 cooperation strategies, 209–211
 diffusion LMS, 211–218
 LMS, 208–222
 steady-state performance, 218–219

Distributed sparsity-promoting algorithms, 483

α -Divergence, 682

Diverging connection, 756

Division algebra, 552

DNA sequences, 314–318

Doubly stochastic matrix, 213–214

D -separation, BNs, 755–758, 758f

Dual frames, 498

Dynamic Bayesian networks, 832–833

Dynamic graphical models, 816–818

E

Echo canceller, 125f

Echolocation signals, time-frequency analysis, 493–497

Eckart-Young-Mirsky theorem, 242

Edgeworth expansion, PDF, 1021–1022

Eigenvalues

covariance matrix, 175f

unequal, 172f, 174f

EKF. *See* Extended Kalman filter (EKF)

Elastic net regularization, 472

EM algorithm. *See* Expectation-maximization (EM) algorithm

Empirical bayes method, 600

Empirical loss functions, 93–94

Energy conservation method, 187

Entropy, 44–45, 302–303

binary random variable, 46f

continuous random variable, 46–47

differential entropy, 47

relative, 47–48

Epigraph, 332, 405–407, 406f

Equality constraints, 1023–1029

Equalizer, 127, 127f

Ergodicity, 31

Ergodic Markov chain Monte Carlo methods, 723–728

Erlang distribution, 27

Error bounds, NORMA, 559–560

Error-correcting codes, 770–772

Error covariance matrix, 141, 150–152

Errors-in-variables regression models, 262

Error vector

convergence, 181–186

covariance matrix, 183–184

Estimation

error, 94, 374–376

interpretation power, 407

nonparametric modeling and, 95–97

Euclidean distance, 283, 285

Euclidean norm, 395, 404–405

descent directions, 166f

graphs, 250f

Euclidean space, 109

Evidence function, 593–595, 596–600

Exact inference methods

chain graph, 773–776

trees, 777–778

Excess mean-square error, 184

Expectation-maximization (EM) algorithm, 598

convergence criterion, 607

description, 606–608

E-step, 607, 623

linear regression, 610–612

lower bound maximization view, 608–610

missing data, 608

Monte Carlo methods, 720–721

M-step, 607, 623

Newton-type searching techniques, 607

online versions, 609

Expectation propagation, 679–683

Expectation step, hidden Markov model, 825–827

Expected loss, 93–94, 177–178

Expected risk, 177–178

Expected value, 15

Explaining away, 756

Exponential distribution, 25, 711

Exponential family

advantage, 600

of probability distributions, 600–606, 644–645

Exponentially weighted isometry property (ERIP), 480

Exponentially weighted least-squares cost function, 245–246

time-iterative computations, 246–247

time updating, 247–248

Extended Kalman filter (EKF), 152, 854

Extreme Learning Machines (ELMs), 900

F

Factor analysis, 972, 977–980

Factor graphs, 768–772

Factorial hidden Markov model (FHMM), 829–832

Factorization

pdf, 643

theorem, 71

Far-end speech signal, 125

Fast iterative shrinkage-thresholding algorithm (FISTA),
459, 461

Fast Newton transversal filter (FNTF) algorithm, 257

Fast proximal gradient splitting algorithm, 386
 FDR. *See* Fisher's discriminant ratio (FDR)
 Feasibility set, 349
 Feasible points, 1025
 Feasible region, 1025
 Feature generation
 phases, 295–296, 295f
 stage, 2
 Feature map, 517
 Feature selection
 phases, 295–296, 295f
 stage, 2
 Feature space, 2, 517
 Feature variable, 60–61
 Feature vector, 2, 60–61
 Feed-forward neural networks, 882–886
 deep learning, 914–915
 hidden layer, 884
 multilayer, 882–886
 output neuron, 884
 universal approximation property, 899–902
 Fill-in edge, 798
 Finite rate of innovation sampling, 436
 First order convexity condition, 331
 Fisher-Neyman factorization theorem, 71
 Fisher's discriminant ratio (FDR), 296–297
 Fisher's information matrix, 1019
 Fisher's linear discriminant, 294–300
 FISTA. *See* Fast iterative shrinkage-thresholding algorithm (FISTA)
 Fixed interval, 866
 Fixed lag smoothing, 866
 Fixed point set, 339
 Focal underdetermined system solver (FOCUSS) algorithm, 472
 Forward-backward algorithm, 827
 Forward-backward splitting algorithms, 385–386
 Forward MSE optimal predictors, 134–138
 Fourier transforms, 33
 analog signal, 435f
 software packages to, 122–123
 Frames theory, 497–502
 Free energy, 608
 Frequency approach, bag-of-words, 571
 Frequentist techniques, 586
 Frobenius norm, 265
 Functional brain networks (FBN), 998
 Functional magnetic resonance imaging (fMRI)
 BOLD contrast, 998–999

functional brain networks, 998
 goals, 999
 ICA, 1000
 scanning procedure, 1000, 1000f
 Function transformation, 711–715

G

Gabor frames, 490–492, 493, 496f
 Gabor transform, 490–492
 Gabor type signal expansion, 414–415
 Gamma distribution, 26–27
 Gating functions, 621
 Gaussian distribution, 183–184, 276
 continuous variables, 20–24
 decision surfaces, 282–287
 hypersurfaces, 282–287
 isovalue contours for, 23f
 multivariate, 21–22, 24
 pdf, 22f, 24
 sub-gaussian distribution, 196–197
 Gaussian-gamma distribution, 603, 603f
 Gaussian-gamma pair, 601–602
 Gaussian Gaussian-gamma pair, 602–603
 Gaussian kernel, 520, 521f, 665f, 688
 Gaussian mixture modeling, 613–620, 651–654
 Gaussian noise case, nonwhite, 84
 Gaussian pdf, 655–656
 computational advantages, 759–760
 conditional, 632–633, 634–637
 joint, 632–634
 marginal pdf, 633–634
 with quadratic form exponent, 631
 Gaussian processes (GP), 687–693
 Gauss-Markov theorem, 143–145, 237–238
 Generalization, 91
 Generalization error, 80–81
 Generalized forward-backward algorithm, 782
 Generalized linear models, 510–511
 Generalized maximum likelihood, 600
 Generalized Rayleigh ratio, 296–297
 Generalized thresholding (GT), 483
 Generative learning, 63–64
 Generic particle filtering, 860–861
 Genes, 315–316
 Geometric averaging rule, 305
 Gibbs distribution, 763
 cliques, 763, 769f
 I-map, 764

- Gibbs sampling, 733–735
 - blocking, 734
 - change-point detection, 738
 - collapsed, 735
 - slice-sampling algorithm, 735
- Gini index, 303
- Givens rotations, 256
- Global decomposition, likelihood function, 836–837
- Gradient averaging, 378
- Gradient descent algorithm, 163, 165, 166*f*, 173*f*
- Gradient descent scheme, 887–894
 - adaptive momentum, 893
 - algorithm, 891–892
 - backpropagation algorithm, 891–892
 - gradient computation, 889
 - iteration-dependent step-size, 893
 - logistic sigmoid neuron, 887
 - momentum term, 893
 - paramount importance, 895
 - pattern-by-pattern/online mode, 892
 - quickprop algorithm, 895
- Gradient vector, 163–165, 165*f*
- Gram-Schmidt orthogonalization, 138, 256
- Graph embedding, 989
- Graphical models
 - dynamic, 816–818
 - for error-correcting codes, 770–772
 - learning structure, 837
 - need for, 746–748
 - parameter estimation, 833–837
 - probabilistic, 751
 - undirected, 762–768
- Graphs
 - bipartite, 769
 - definitions, 749–753
 - direction/undirected, 749
 - factor, 768–772
 - triangulated, 796–804
 - undirected (*see* Undirected graph)
- Graph theory, 746
- Greedy algorithms, 451–456
 - CSMP, 455–456, 460
 - LARS, 454
 - OMP, 451, 453
- H**
- Halfspace, 347–348
- Hamiltonian Monte Carlo methods, 736
- Hammerstein model, 511–514
- Hard thresholding
 - function, 456–457, 459, 460
 - operation, 409–411, 410*f*
- Head-to-head connection, 756
- Head-to-tail connection, 755
- Heat bath algorithm, 733
- Heavy-tailed distribution, 671
- Hermitian operation, 16–17
- Hessian matrix, 292, 293–294, 377–378
- Hidden Markov model (HMM), 816, 817–818
 - autoregressive, 828
 - discriminative learning, 828–829
 - expectation step, 825–827
 - FHMM, 829–832
 - inference, 821–825
 - left-to-right type, 819, 820*f*
 - maximization step, 827
 - parameters, 821, 825–828
 - sum product algorithm, 821
 - time-varying dynamic Bayesian networks, 832–833
 - transition probability, 818, 819
 - variable duration, 829
 - Viterbi reestimation, 827–828
- Hidden variables, 606
- Hierarchical Bayesian modeling, 647, 695–696
- Hierarchical mixture of expert (HME), 625, 626*f*
- Hierarchical priors, 599
- High-definition television (HDTV) system, 412–413
- Hilbert space, 329, 397
 - closed convex set, 334, 337, 339–340, 344
 - convex set, 334
- Hinge loss function, 348–349, 349*f*, 538–539, 558–559
- Histogram technique, 95–96
- HME. *See* Hierarchical mixture of expert (HME)
- HMM. *See* Hidden Markov model (HMM)
- Homotopy algorithm, 454
- Householder reflections, 256
- Huber loss function, 530–531, 531*f*
- Hyperparameters, 599, 600
 - Gaussian processes, 691–692
 - support vector machine, 550–551
- Hyperplane, 60, 61–62
- Hyperprior, 647
- Hyper rectangles, 300–301
- Hyperslab, 345–346
- Hyperspectral image unmixing (HSI), 693–699
 - experimental results, 696–699
 - hierarchical Bayesian modeling, 695–696

Hyperspectral remote sensing, 693–694

Hypersurfaces, 280–281, 282

Gaussian distribution, 282–287

naïve Bayes classifier, 287–288

nearest neighbor rule, 288–290

Hypothesis class, 371

Hypothesis space, 528–529

I

IIR. *See* Infinite impulse response (IIR)

Ill conditioning, 74–76

Image deblurring, 121–124

I-maps

BNs, 761–762

Markov Random Fields, 763–765

Importance sampling (IS), 718–720

Impulse response function, 411–412, 412*f*

Incremental networks, 210

Incremental topology, 211*f*

Independence assumption, 182

Independent component analysis (ICA), 944

ambiguities, 958

cocktail party problem, 963–966

Edgeworth expansion, 959–960

fourth-order cumulants, 957–958

Gaussian distributions, 956

gradient ascent scheme, 960–961

Infomax principle, 962

Kullback-Leibler divergence, 959–960

maximum likelihood, 962

mixture variables, 955

mutual information, 959–960

natural gradient, 961–962

negentropy, 963

non-Gaussian distributions, 958–959

Riemannian metric tensor, 961–962

tensorial methods, 958

unmixing/separating matrix, 955–956

Inequality constraints, 1025–1029

Inference, 684, 821–825

Infinite impulse response (IIR), 120–121

Information filtering scheme, 152

Information projection (I-projection), 679

Information theory, 41

continuous random variables, 45–48

discrete random variables, 42–45

Inner product space, 395

Innovations process, 152

Input space, 2

Input vector, 57–58

Intercausal reasoning, 756

Intercept, 57–58

Interference cancellation, 124–125

Interior point methods, 358–359

Interpretation power of estimator, 407

Intersymbol interference (ISI), 127, 412–413

Intrinsic dimensionality, 90–91, 938, 938*f*, 939

Invariant distribution, 722

Inverse Fourier transform, 35

Inverse problems, 74–76

Inverse system identification, 126

Invertible transformation, 17, 667

IRLS. *See* Iterative reweighted Least Squares scheme (IRLS)

IS. *See* Importance sampling (IS)

ISI. *See* Intersymbol interference (ISI)

Ising model, 765–767

Isodata algorithm, 618

Isometric mapping (ISOMAP), 987–991

IST algorithms. *See* Iterative shrinkage/thresholding (IST) algorithms

Iterative hard thresholding (IHT), 466–467, 466*f*

Iterative refinement algorithm, 383–384

Iterative reweighted Least Squares scheme (IRLS), 293, 471–472

Iterative shrinkage/thresholding (IST) algorithms, 456–462

Iterative soft thresholding (IST) algorithms, 466–467, 466*f*

J

Jacobian matrix, of transformation, 17–18

Joint distribution, 748, 749

Joint Gaussian pdf, 632–634

Jointly distributed random variables, 77

Jointly sufficient statistics, 71

Joint pdf, 68, 71, 836–837

Joint probabilities, 12–13

Join tree, construction, 799–801

Junction tree, 798, 801–804

K

Kalman filtering, 149, 851–854, 853*f*

Kalman gain, 150–152, 246–247, 248, 257

Karush-Kuhn-Tucker (KKT) conditions, 1025–1026

Kernel APSM (KAPSM) algorithm, 560–565

classification, 561–565

nonlinear equalization, 564–565

quantized, 562–563, 565

regression, 560–561

Kernel Hilbert spaces, 152
 Kernel LMS (KLMS), 553–556
 Kernel perceptron algorithm, 881–882
 Kernels, 96–97
 construction, 523–524
 covariance functions, 688–689
 function, 520–525
 matrix, 519, 688–689
 ridge regression, 528–530, 537–538
 trick, 517, 532, 537–538
 Kikuchi energy, 815–816
 k -means algorithm, 618, 619
 k -nearest neighbor density estimation, 97
 k -nearest neighbor (k -NN) rule, 288–290
 k -rank matrix approximation, 242
 KRLS, 565
 k -spectrum kernel, 525
 Kullback-Leibler distance, 305
 Kullback-Leibler (KL) divergence, 47–48
 EM algorithm, 608–609, 610f
 mean field approximation, 642, 643
 minimizing, 680–681
 Kurtosis, 958, 1020–1021

L

Labeled faces in the wild (LFW) database, 947
 Lagrange multipliers, 1024–1025
 Lagrangian, 205
 duality, 1027–1028
 function, 1024–1025
 Laplacian approximation, 662, 664
 evidence function, 596–600
 method, 596–600
 Laplacian kernel, 521
 Laplacian pdf, 668–670, 670f, 671, 672f
 Large scale tasks, 376
 LARS-LASSO algorithm, 454, 462
 Latent Markov model, 816–817, 818f
 Latent variables, 606–610
 Lattice-ladder algorithm, 132
 forward/backward MSE optimal predictors, 134–138
 orthogonality of optimal backward errors, 138–140
 Toeplitz matrix, 133
 LDA. *See* Linear discriminant analysis (LDA)
 Learning, 1
 curve, 171–174
 from data, 1
 deep (*see* Deep learning)

 sparsity-aware (*see* Sparsity-aware learning)
 Least absolute shrinkage and selection operator (LASSO), 407–411
 adaptive norm-weighted, 477–478
 asymptotic performance, 475–477
 elastic net regularization, 472
 group, 467–468
 LARS algorithm, 454
 regularized cost function, 458
 Least angle regression (LARS) algorithm, 454, 466–467
 Least mean fourth (LMF) algorithm, 196–197
 Least-mean-square (LMS)
 adaptive algorithm, 179–188
 algorithm, 179–180, 368
 consensus matrix, 223–224
 convergence, 181–186, 199f, 200f, 201f
 cumulative loss bounds, 186–188
 diffusion, 211–218
 distributed learning, 208–222
 H^∞ optimality of, 187
 linearly constrained, 204–206
 normalized, 193–194
 parameter estimation, 209
 recursion, 213
 relatives of, 196
 sign-error, 196
 steady-state performance, 181–186, 206
 target localization, 222–223
 time-varying model, 206–207
 tracking performance, 206–208
 transform-domain, 197–201
 widely-linear, 195
 Least modulus method, 530–531
 Least-squares (LS) estimator
 asymptotic distribution of, 238–239
 BLUE, 237–238
 covariance matrix, 236–237
 Cramer-Rao bound, 238
 loss criterion, 276, 308f, 311
 unbiased, 236
 Least-squares method
 classifier, 61–62
 computational aspects, 255–257
 fitting plane, 60f
 linear classifier, 63f
 linear regression, 234–236, 235f
 loss function, 56–57, 58–59, 59f
 minimization task, 72–73
 optimal, 59

- regularization, 72–73
 - ridge regression, 243–245
 - unregularization, 72–73
 - Leave-one-out (LOO) cross-validation method, 92–93
 - Levenberg-Marquardt method, 376–377
 - Levinson algorithm, 132–140
 - Levinson-Durbin algorithm, 137
 - Likelihood function, 82
 - Linear classifier, 63*f*, 283
 - Linear congruential generator, 709–710
 - Linear convergence, 167
 - Linear discriminant analysis (LDA), 286, 291
 - Linear discriminant, Fisher's, 294–300
 - Linear dynamical systems (LDS), 817–818
 - Linear filtering, 35–36, 118–120
 - Linear Gaussian models, 759–760
 - Linear ϵ -insensitive loss function, 346, 347*f*, 530–537, 559
 - Linear independency, 394
 - Linear inverse problems, 438
 - Linear kernel, 688
 - Linearly constrained LMS, 204–206
 - Linearly constrained minimum variance (LMV), 148
 - Linearly separable classes, 515–517
 - classes, 540–545
 - probability, 515*f*
 - two-dimensional plane, 516*f*
 - Linear regression, 57–60
 - Bayesian approach, 589–593
 - dependencies, 646*f*
 - EM algorithm, 610–612
 - MAP estimator, 588–589
 - ML estimator, 587
 - nonwhite Gaussian noise case, 84
 - variational Bayesian approach to, 645–651
 - Linear space, 393
 - Linear time invariant (LTI), 35–36, 512–514
 - Linear varieties, 343, 343*f*
 - LMF algorithm. *See* Least mean fourth (LMF) algorithm
 - LMS. *See* Least-mean-square (LMS)
 - LMV. *See* Linearly constrained minimum variance (LMV)
 - ℓ_0 norm minimizer, 417–418
 - equivalence, 426–429
 - uniqueness, 422–426
 - ℓ_1 norm minimizer, 418
 - characterization, 419
 - equivalence, 426–429
 - ℓ_2 norm minimizer, 416*f*, 417
 - Local independencies, 749–750
 - Local linear embedding (LLE), 986–987
 - Log-concave function, 805–807
 - Log-convex function, 808
 - Logistic regression, 290–294, 662–666
 - Logistic sigmoid function, 290–291, 662, 887
 - Log-likelihood function, 82–83, 292
 - Log-loss function, 311–313
 - Log-odds ratio, 311
 - Log-partition, 815
 - Loopy belief propagation, 813–816
 - Loss functions
 - empirical, 93–94
 - expected, 93–94
 - mean-square-error (*see* Mean-square-error (MSE) loss function)
 - optimizing, 106
 - parametric modeling, 56
 - Loss matrix, 279–280
 - Lower bound, computation, 650–651
 - Low-rank matrix factorization method
 - matrix completion, 991–994
 - robust PCA, 995–996
 - LTI. *See* Linear time invariant (LTI)
 - LTIFIR filter, 137
- ## M
- Magnetic resonance imaging (MRI), sparsity-promoting learning, 473–474
 - Mahalanobis distance, 283, 285
 - Majority voting rule, 306
 - Majorization-minimization techniques, 458, 471
 - Manifold learning, 434
 - MAP estimator. *See* Maximum a posteriori probability (MAP) estimator
 - Marginal pdf, 633–634
 - Marginal probabilities, 13, 849
 - Markov blanket, 758
 - Markov chain Monte Carlo methods
 - auxiliary variable, 735
 - building, 724
 - detailed balanced condition, 723
 - ergodic, 723–728
 - invariant distribution, 722
 - reversible jump, 736
 - transition probabilities matrix, 721–722
 - Markov condition
 - causality, 753–755
 - completeness, 761–762
 - definitions, 749
 - d -separation, 755–758, 758*f*

- Markov condition (*Continued*)
 - faithfulness, 761–762
 - graphs, 749–753
 - I-maps, 761–762
 - linear Gaussian models, 759–760
 - multiple-cause networks, 760, 761f
 - soundness, 761–762
- Markov networks, 762
- Markov Random Fields (MRF), 762
 - Boltzmann machine, 809f
 - I-maps, 763–765
 - independencies, 763–765
 - Ising model, 765–767
- MARTs. *See* Multiple additive regression trees (MARTs)
- Matching Pursuit, 451
 - CoSaMP, 455–456, 480
 - CSMP algorithms, 455–456, 460, 479–480
 - OMP, 451, 453, 466–467
- Matrices
 - derivatives, 1014
 - inversion lemmas, 1014
 - positive definite and symmetric, 1015
 - properties, 1013–1014
- Matrix completion, 991–994
 - applications, 997
 - collaborative filtering task, 996–997
- Maximal cliques, 763, 764f, 768
- Maximal spanning tree algorithm, 799
- Maximum a posteriori probability (MAP) estimator, 88–89, 588–589, 592
- Maximum entropy (ME) method, 47, 605–606
- Maximum likelihood (ML) method, 82–84, 82f, 277
 - estimator, 587
 - Type II, 600
- Maximum margin classifiers, 540–545
- Maximum variance unfolding method, 989
- Max-product algorithms, 782–789
- Max-sum algorithms, 782–789
- McCulloch-Pitts neuron, 880–881
- MDA. *See* Mirror descent algorithms (MDA)
- Mean, 15–17
- Mean field approximation, 641–645, 810–813
- Mean field equation, 811
- Mean field factorization, 810
- Mean square deviation (MSD), 358, 359f
- Mean-square-error (MSE), 65
 - cost function, 176
 - curves, 260f, 262f
 - estimation, 77–78
 - iteration functions, 565, 566f
 - linear estimator, 178–179
 - local cost function, 212–213
 - values, 76, 76t
- Mean-square error linear estimation, 105–106, 141–148
 - complex-valued variables, 111–118
 - constrained linear estimation, 145–148
 - cost function, 107–108, 108f
 - deconvolution, 121–124, 126–132
 - Gauss-Markov theorem, 143–145
 - geometric viewpoint, 109–111
 - interference cancellation, 124–125
 - Kalman filtering, 149
 - Lattice-ladder algorithm, 132–140
 - Levinson algorithm, 132–140
 - linear filtering, 118–120, 119f, 120–124
 - minimum, 110f
 - normal equations, 106–108
 - optimal equalizer, 130–131
 - system identification, 125–126
- Mean-square-error (MSE) loss function
 - complex-valued case, 175–176
 - cost function, 167
 - cross-correlation vector, 171–174
 - error curves, 171–174
 - gradient descent algorithm, 173f
 - learning curve, 171–174
 - minimum eigenvalue, 169–171
 - parameter error vector convergence, 171
 - time constant, 169
 - time-varying step sizes, 174–176
- Mean-square sense, convergence in, 49
- Measurement noise, 149–150
- Mercedes Benz (MB) frame, 499–500
- Message-passing algorithms, 460–462
 - exact inference methods, 773–789
 - junction tree, 801–804
 - max-product algorithms, 782–789
 - max-sum algorithms, 782–789
 - sum-product algorithm, 778–782
 - two-way, 801–803
- Metropolis-Hastings algorithm, 729, 730, 735, 736
- Metropolis method, 728–729
 - burn-in phase, 730
 - convergence issues, 731–732
- MIMO systems. *See* Multiple-input-multiple-output (MIMO) systems
- Minimum distance classifiers, 285–287

Minimum variance distortionless response (MVDR)
 beamforming, 148

Minimum variance unbiased estimator (MVUE), 66–67, 144–145, 238

Min-Max duality, 1026–1027

Mirror descent algorithms (MDA), 388–389

Misclassification error, Bayesian classification, 277–280

Mixing linear regression models, 622–625
 HME, 625, 626f
 mixture of experts, 624–625

Mixing logistic regression models, 625–627

Mixing of learners, 621

Mixing time, 730

Mixture of experts, 621, 621f, 624–625

Mixture of factor analyzers (MSA), 978–979

Mixture scatter matrix, 296

ML method. *See* Maximum likelihood (ML) method

Mode, 169–171, 170f

Model-based Compressed Sensing, 468–469

Modulated wideband converter (MWC), 435–436

Moment generating function, 1020

Moment matching, 680–681, 682

Moment projection (M-projection), 680

Moments, 1020–1021

Monte Carlo methods, 709
 advantages, 736–737
 change-point detection, 737–738
 concepts, 708–710
 EM algorithm, 720–721
 Gibbs sampling, 733–735
 Hamiltonian function, 736
 importance sampling, 718–720
 Markov chain, 721–728
 Metropolis method, 728–732
 random sampling, 711–715
 rejection sampling, 715–718

Moore-Penrose pseudo-inverse, 234–236

Moreau envelop, 379, 381f, 460

Moreau-Yosida regularization, 379

Moving average model, 40

Multichannel estimation, 112–113, 141

Multiclass Fisher's discriminant, 299–300

Multiclass generalizations, SVM, 552–553

Multidimensional Scaling (MDS), 946

Multinomial distribution, 19–20

Multinomial resampling, 847

Multiple additive regression trees (MARTs), 314

Multiple-cause Bayesian networks, 760, 761f, 805–807

Multiple-input-multiple-output (MIMO) systems, 412–413

Multiple kernel learning (MKL), 567–568

Multiple measurement vectors (MMV), 659

Multipulse signals, 436

Multitask learning, 548

Multivariate Gaussian distribution, 21–22, 24

Multivariate linear regression (MLR), 955

Mutual coherence, 424–426

Mutual information, 43–44

MVDR beamforming. *See* Minimum variance distortionless response (MVDR) beamforming

MVUE. *See* Minimum variance unbiased estimator (MVUE)

MWC. *See* Modulated wideband converter (MWC)

N

Naive Bayes classifier, 287–288

Naive online R_{reg} minimization algorithm (NORMA), 556–560

Natural gradient, 376–377

Natural parameters, 600

Near-end speech signal, 125

Nearest neighbor rule, 288–290

Near-to-Toeplitz, 256–257

NESTA algorithm, 461, 472–473, 490, 495–496

Neural networks
 backpropagation algorithm, 886–897
 Bayesian learning, 902–903
 feed-forward, 882–886
 gradient descent scheme, 887–894
 perceptron algorithm, 876
 pruning, 897–899
 synapses, 876

Newton's iterative minimization method, 248–251

Newton's scheme, 293

NLMS. *See* Normalized least mean square (NLMS)

Noise cancellation, 127, 127f

Noisy-OR model, 746–747, 805–807

Nonempty set, 683

Noninformative/objective priors, 599

Nonlinear dimensionality reduction
 ISOMAP, 987–991
 kernel PCA, 980–982
 Laplacian eigenmaps, 982–986
 local linear embedding, 986–987

Nonlinear filter, 512f

Nonlinear manifold learning, 979

Nonnegative garrote, 411, 412f

Nonnegative matrix factorization (NMF), 971–972

Non-negative real function, 38

Nonoverlapping training sets, 93

Nonparametric Bayesian modeling, 54, 95–97, 683–686
 estimation, 95–97
 representer theorem, 528

Nonparametric sparsity-aware learning, 568–570

Nonseparable classes, SVM, 545–548

Nonsmooth convex cost functions
 linearly separable, 364
 minimizing, 362–367
 optimizing, 358–370
 subdifferentials, 359–362
 subgradients, 359–362, 363–365

Nonstationary environments, LMS, 206–208

Nonwhite Gaussian noise, 84

Norm, 395
 definition, 404–405
 ℓ_0 minimizer, 417–418, 422–429
 ℓ_1 minimizer, 418, 419, 426–429
 ℓ_2 minimizer, 416f, 417
 searching for, 404–407

Normal distribution, 20–24

Normal equations, 110

Normal factor graph (NFG), 770

Normalized graph Laplacian matrix, 984–985

Normalized least mean square (NLMS)
 convex analytic path, 353
 stochastic gradient descent, 193–194, 201, 202f

Normed linear space, 395

Nucleobases, 314–315

Nucleotides, 314–316

O

OBCT. *See* Ordinary binary classification trees (OBCT)

Observations, 57–58

Occam's razor rule, 78–79, 593–600, 643

OCR systems. *See* Optical character recognition (OCR) systems

OMP. *See* Orthogonal matching pursuit (OMP)

One-against-all, 552

One-against-one, 552

One pixel camera, 432–433

Online cyclic coordinate descent time weighted LASSO (OCCD-TWL), 478, 484f

Online learning
 approximation error, 374–376
 batch vs., 376–379
 and big data applications, 374–379
 convex, 367–370
 estimation error, 374–376
 expected loss/risk function, 374

optimization error, 374–376
 techniques, 162

Online perceptron algorithm, 880

Optical character recognition (OCR) systems, 2, 923

Optimal brain damage technique, 898

Optimal brain surgeon method, 898

Optimal linear estimation, 124

Optimization error, 374–376

Order statistics, 30

Ordinary binary classification trees (OBCT), 300–301, 300f, 304

Ordinary-differential-equation approach (ODE), 187

Ornstein-Uhlenbeck kernel, 689

Orthogonality
 geometric viewpoint, 109–111
 optimal backward errors, 138–140

Orthogonal matching pursuit (OMP), 451
 algorithm, 466–467
 recover optimal sparse solutions, 453

Orthogonal projection, 109, 110

Outlier, 537

Output variable, 57–58

Overcomplete dictionaries, 414–415

Overdetermined system, 240–242

Overfitting, 74–76

P

Pairwise MRFs undirected graphs, 766, 767f

Parallelogram law, 396

Parameter error vector convergence, 171

Parametric functional form, 54–55

Parametric modeling, 53
 curve fitting problem, 54–55, 55f
 deterministic point of view, 54–57
 loss function, 56
 nonnegative function, 56

Parity-check bits, 770–771

Parseval tight frame, 498–499

Partial least-squares (PLS) method, 954–955

Particle filtering, 854
 auxiliary, 862–868
 degeneracy phenomenon, 858–860
 generic, 860–861
 one-dimensional random walk model, 857–858
 SIS, 855, 856
 state-space model, 853f, 854–855

Parzen windows, 96–97

Path, 749

- Pattern, 60–61
- Pattern recognition, 60–61, 91, 295–296
- Peak signal-to-noise ratio (PSNR), 438–439
- Perceptron algorithm, 364–365, 876, 878
- Perceptron cost, 877–882
- Perfect elimination sequence, 798
- Perron-Frobenius theorem, 722
- Persistent contrastive divergence (PCD) algorithm, 913, 914
- PGM. *See* Projected gradient method (PGM)
- pmf. *See* Probability mass function (pmf)
- POCS. *See* Projections onto convex sets (POCS)
- Poisson process, 737–738
- Polynomial kernel
 - homogeneous, 520
 - inhomogeneous, 520, 522f
- Population-based methods, 735–736
- Positive definite, 16, 34, 1015
- Positive definite kernel, 519, 520
- Positive semidefinite, 1015
- Posteriori probability, 63–64, 276
- Potential functions, 96–97, 762
- Potts model, 766
- Power spectral density (PSD), 33–38, 120–121
 - definition, 35
 - physical interpretation of, 37–38
- Prediction, 118, 186
- Preprocessing stage, 32–33
- Primal estimated subgradient solver for SVM (PEGASOS)
 - algorithm, 369–370, 551
- Principal axes/directions, 244–245
- Principal component pursuit (PCP), 995
- Principal components regression, 244–245
- Principia Mathematica, 754–755
- Principle component analysis (PCA)
 - eigenimages/eigenfaces, 947, 947f, 948f
 - feature generation, 943–944
 - latent semantics indexing, 943
 - latent variables, 944–949
 - LFW database, 947
 - low-rank matrix factorization method, 942
 - minimum error interpretation, 943
 - mutually uncorrelated, 943–944
 - online subspace tracking, 949, 950f
 - optimization task, 940–941
 - principle directions, 940
 - supervised PCA, 947
 - SVD decomposition, 941, 942
- Probabilistic PCA (PPCA), 974–977
- Probability density function (pdf), 10
 - beta distribution, 26f
 - definition, 18f
 - edgeworth expansion, 1021–1022
 - gamma distribution, 27f
 - Gaussian, 22f, 24
 - uniform distribution, 21f
- Probability distributions
 - exponential family, 600–606, 644–645
 - random walk chain, 726–728, 727f, 728f
- Probability mass function (pmf), 12, 19f
- Probit regression, 294
- Process noise, 149–150
- Product rule of probability, 13
- Projected gradient method (PGM), 365–366
- Projected Landweber method, 366
- Projected subgradient method, 366–367
- Projection approximation subspace tracking (PAST), 949
- Projections onto convex sets (POCS)
 - algorithm, 344
 - analytical expressions, 335–336
 - classification, 347–349
 - concepts, 333–335
 - fundamental theorem, 341–344
 - halfspace, 336f
 - hyperplane, 336f
 - intersection, 345–346
 - linear varieties, 343, 343f
 - nonempty intersection, 341, 342
 - nonexpansiveness property, 340f, 343–344
 - parallel version, 344
 - product spaces, 344
 - properties, 337–341
 - regression, 345–347
 - relaxed, 339–340, 340f, 341f
 - weak convergence, 342
- Property sets, 349
- Proportionate NLMS, 194
- Protein folding prediction, 314–318
- Proteinogenic amino-acids, 314–315
- Proximal forward-backward splitting operator, 386–387
- Proximal gradient splitting algorithms, 385–386
- Proximal mapping, 460
- Proximal operators, 379–385
 - minimization, 383–385
 - properties, 382–383
 - splitting methods, 385–389
 - subdifferential mapping, 384–385

Pruning, neural networks

- convolutional networks, 899
- early stopping, 898–899
- optimal brain damage technique, 898
- weight decay, 897
- weight elimination, 897
- weight sharing, 898

Pruning tree, 303–304

PSD. *See* Power spectral density (PSD)

Pseudo covariance, 114–115

Pseudo-inverse matrix, 240–242

Pseudorandom generator, 709–710, 711

Q

QR factorization, 255–256

Quadratic discriminant analysis (QDA), 286

Quadratic form exponent, pdfs with, 631

Quadratic ϵ -insensitive loss function, 530–531, 531f, 536

Quantized KLMS (QKLMS), 555–556, 565

Quasi-stationary process, 818

Quickprop algorithm, 895

R

Random demodulator (RD), 432, 434–435, 436

Random field, 121

Random forests, 303–304

Random-modulation pre-integrator (RMPI), 434–435

Random number generation, 709–710

Random sampling, 711–715

Random signal, 29

Random variables

- axiomatic definition, 11–12
- complex, 16–17
- continuous (*see* Continuous random variables)
- discrete (*see* Discrete random variables)
- geometric interpretation of, 109–111
- probability and, 10–18
- relative frequency definition, 11
- transformation of, 17–18

Random vector, 15

Rao-Blackwellization technique, 866

Rao-Blackwell theorem, 70, 71, 735

Rate of convergence, 457–458

Rational quadratic kernel, 689

Rayleigh fading channel, 342

RD. *See* Random demodulator (RD)

Real linear space, 394

Recursive least-squares (RLS) algorithm, 234, 245–248

convergence curve, 354

fast versions, 256–257

Newton's method, 251

simulation examples, 259, 260–261

steady state performance, 252–254

widely linear, 254–255

Reduced convex hull interpretation, 548

Regression, 3–8

Bayesian, 690–692

deming, 262

errors-in-variables, 262

input-output relation, 57f

KAPSM algorithm, 560–561

least-squares linear, 234–236, 235f

linear, 57–60

linear ϵ -insensitive loss function, 559

POCS, 345–347

principal components, 244–245

ridge, 243–245

Regressor, 57–58

Regret analysis, 367–368, 370–374

Regularity assumption, 1023

Regularization, 72–76

Regularized dual averaging (ARD) algorithm, 388

Regularized particle filter, 866

Rejection sampling, 715–718

Relative entropy, 47–48, 896–897

Relevance vector machine (RVM), 661–666

Relevance vectors, 662

Representer theorem, 525–528

nonparametric, 528

semiparametric, 527

Reproducing kernel Hilbert spaces (RKHS), 95, 517–518, 662

authorship identification, 570–573

definition, 510

generalized linear models, 510–511

KAPSM algorithm, 560–565

kernel functions, 520–525

kernel LMS, 553–556

kernel trick, 517

NORMA, 556–560

properties, 519–520

representer theorem, 525–528

ridge regression, 528–530, 537–538

theoretical highlights, 519–520

Resampling, 847–849, 851

Restricted Boltzmann machine (RBM), 905–906, 908–914

Restricted isometry property (RIP), 427–429

Reversible jump Markov chain Monte Carlo algorithms, 736

Ridge regression, 72–73, 243–245
 kernels, 528–530, 537–538
 principal components regression, 244–245
 Right stochastic matrix, 213–214
 Ring networks, 210
 Ring topology, 211*f*
 RIP. *See* Restricted isometry property (RIP)
 RKHS. *See* Reproducing kernel Hilbert spaces (RKHS)
 RLS algorithm. *See* Recursive least-squares (RLS) algorithm
 Robbins-Monro algorithm, 177–179
 Robust loss functions, 311–313
 Robust PCA
 applications of, 997
 low-rank matrix factorization, 995–996
 Robust sparse signal recovery, 429–430
 Running intersection property, 798

S

Saddle point condition, 1027
 Saliencies, 898
 Sample mean, 31
 Sample sequences, 29
 Sample space, 12
 Sampling-importance-resampling (SIR), 860–861
 SCAD. *See* Smoothly clipped absolute deviation (SCAD)
 Scatter matrices, 295–296
 Schur algorithm, 140
 Schur complement, 133–134
 Search direction, 163
 Second order convexity condition, 331
 Segmental k -means training algorithm, 827–828
 Semiparametric representer theorem, 527
 Semisupervised learning, 3, 64, 202–203
 Separator, 799, 801–804
 Separator nodes, 802–803
 Sequential importance sampling (SIS), 845–846, 850
 importance sampling revisited, 846–847
 particle filtering, 855, 856
 resampling, 847–849, 851
 sequential sampling, 849–851
 Serial connection, 755
 Set-membership algorithms, 354
 Shepp-Logan image phantom, 474*f*
 Shrinkage methods, 314
 Sigmoidal Bayesian networks, 758–759, 759*f*, 760
 Sigmoid link function, 290, 291*f*
 Signal compression, 412–413
 Signal processing, filtering, 118

Signal restoration, 438
 Sign-error LMS, 196
 Sinc kernel, 523
 Single-layered feed-forward networks (SLFNs), 900
 Single-stage auxiliary particle filter, 865–867
 Singular value decomposition (SVD), 239–242, 941, 942
 SIS. *See* Sequential importance sampling (SIS)
 Slab method, 660–661
 Slack variables, 532
 SLDS. *See* Switching linear dynamic systems (SLDS)
 Slice-sampling algorithm, 735
 Small scale tasks, 376
 Smoothing, 118, 852, 866
 Smoothly clipped absolute deviation (SCAD), 411, 412*f*, 478
 Softmax activation function, 624, 896–897
 Soft thresholding, 380–381
 function, 456–457
 operation, 409–411, 410*f*
 Soundness, 761
 Sparse adaptive projection subgradient method (SpAPSM),
 480–484, 481*f*, 484*f*
 Sparse analysis representation, 485–486
 Sparse Bayesian Learning (SBL), 657–660
 Sparse factor analysis, 977
 Sparse modeling, 404
 Sparse reconstruction by separable approximation (SpaRSA)
 algorithm, 458–459
 Sparse signal representation, 411–415, 487–488
 Sparse solutions, 453, 475
 Sparsity-aware learning, 385, 404
 Bayesian approach to, 655–661
 concave, 675–676
 cost function, 675–676
 Cramer-Rao bound, 679
 de-noising, 438–439
 geometric interpretation, 419–422
 least absolute shrinkage/selection operator, 407–411
 ℓ_0 norm minimizer, 417–418, 422–429
 ℓ_1 norm minimizer, 418, 419, 426–429
 ℓ_2 norm minimizer, 416*f*, 417
 models, 485–490
 nondecreasing, 675–676
 parameter identifiability, 678–679
 robust sparse signal recovery, 429–430
 searching for norm, 404–407
 techniques, 404
 variational parameters, 677
 variations on, 467–474
 Sparsity-aware regression, 671–675

- Sparsity-promoting algorithms, 356, 450
 - adaptive norm-weighted LASSO, 477–478
 - AdCoSaMP algorithm, 479–480
 - distributed, 483
 - frames theory, 497–502
 - greedy algorithms, 451–456
 - iterative shrinkage/thresholding algorithms, 456–462
 - LASSO, 475–477
 - magnetic resonance imaging, 473–474
 - phase transition behavior, 464–465, 465f
 - practical hints, 462–467
 - SpAPSM, 480–484
- Spectral signature, 694
- Spectral unmixing (SU), 694–695
- Spike method, 660–661
- Spline kernels, 521
- Split Levinson algorithm, 137
- Splitting criterion, 300–301, 302–303
- Squared-error loss function, 234
- Squared exponential kernel, 688
- Squashing function, 887
- SSS. *See* Strict-sense stationarity (SSS)
- Stable embedding, 433–434
- Stacking, 306
- State equation, 149–150
- State-observation models, 816–817
- State-space models, 149–150, 816–817
 - Kalman filters, 853f
 - particle filters, 853f, 854–855
- Stationarity, 30
- Stationary iterative/iterative relaxation methods, 456–457
- Statistical filtering, 119f
- Statistical independence, 13
- Statistical signal processing, 5
- Steady-state performance, 218–219
 - distributed learning, 218–219
 - improving, 219
 - LMS in stationary environments, 181–186
 - of RLS, 252–254
- Steepest descent method, 163–167, 293
- Stick-breaking construction, 685–686
- Stochastic approximation, 177–179, 251
- Stochastic convergence
 - almost everywhere, 49
 - distribution, 49–51
 - everywhere, 48
 - mean square sense, 49
 - probability, 49
- Stochastic EM, 720–721
- Stochastic gradient descent schemes, 178–179
- Stochastic processes, 29
 - autoregressive models, 38–40
 - first/second order statistics, 30
 - power spectral density, 33–38
 - stationarity/ergodicity, 30–33
- Stochastic volatility model, 867
- Stop-splitting rule, 303
- Strict-sense stationarity (SSS), 30, 31
- String kernels, 525
- Strongly convex auxiliary function, 388
- Structured sparsity, 467, 468–469
- Subband adaptive filters, 198
- Subdifferential mapping
 - proximal operators, 384–385
 - resolvent of, 384–385
- Sub-Gaussian distribution, 196–197
- Subgradient algorithm, 359–362, 363–365
 - generic scheme, 365
 - regret analysis of, 372–374
- Subjective priors, 599
- Sublinear global rate of convergence, 457–458
- Sub-Nyquist sampling
 - analog-to-information conversion, 434–436
 - definition, 434–435
- Sufficient statistics, 70–72
- Sum-product algorithm, 778–782, 821
- Supervised learning, 3, 64
- Support vector machine (SVM), 369, 538–539, 665
 - applications, 550
 - division and clifford algebras, 552
 - hyperparameters, 550–551
 - linearly separable classes, 540–545
 - multiclass generalizations, 552–553
 - nonseparable classes, 545–548
 - one-against-all, 552
 - one-against-one, 552
 - PEGASOS, 551
 - performance, 550
- Support vector regression (SVR), 662
 - linear ϵ -insensitive loss function, 530–537, 559
 - optimization task, 530–531
- Support vectors, 533, 543
- Switching linear dynamic systems (SLDS), 832
- Synapses, 876
- Systematic resampling, 848
- System identification, 125–126

T

Tail-to-tail connection, 756
 Test error, 80–81
 Thinning process, 731
 Tight frames, 498–499
 Time-adaptive algorithm, 162
 Time-and-norm-weighted LASSO (TNWL), 477–478
 Time constant, 169, 185–186
 Time-frequency analysis
 echolocation signals, 493–497
 Gabor frames, 490–492, 493
 Gabor transform, 490–492
 time-frequency resolution, 492–493
 Time-frequency resolution, 492–493
 Time sequential nature, 140
 Time-shifted versions, 132
 Time-shift structure, 256–257
 Time varying signal, 483–484
 Time varying statistics, 149
 Time-varying step sizes, 174–176
 TNWL. *See* Time-and-norm-weighted LASSO (TNWL)
 Toeplitz matrix, 40, 133
 Total-least-squares (TLS) method, 261–268
 Training data set, 57–58
 Training deep networks
 backpropagation, 907
 distributive representation, 907
 feed-forward networks, 914–915
 restricted Boltzmann machine, 905–906, 908–914
 sparsity, 907
 Training error, 80–81, 92^f
 Training set, 2
 Transform-domain LMS, 197–201
 Transition probability matrix, 722–723, 724, 725
 detailed balanced condition, 723
 hidden Markov model, 818, 819
 Markov chains, 724, 725–726
 properties, 722–723
 Transversal implementation, LTIFIR filter, 137
 Tree reweighted belief propagation, 815
 Trees
 boosting, 313–314
 classification, 300–304, 301^f
 exact inference methods, 777–778
 Triangle inequality, 395
 Triangulated graphs, 796–804
 Bayesian network, 800–801, 800^f
 undirected graph, 796, 797^f, 799

Two-stage-thresholding (TST) algorithms, 460, 466–467, 466^f
 Tychonoff-Phillips regularization, 72
 Type I estimator, 592
 Type II maximum likelihood, 600

U

Unbiased estimation, 31, 65–67
 Underdetermined system, 240–242
 Undirected graph
 perfect elimination sequence, 798
 triangulated graph, 796, 797^f, 799
 Undirected graphical models, 762–768
 CRFs, 767–768
 independencies/I-maps in Markov random fields, 763–765
 Ising model, 765–767
 Uniform distribution, 20, 712^f
 Union of subspaces, 433
 Unit vector, 193^f
 Unobserved random variable, 57–58
 Unscented Kalman filters, 152
 Unsupervised learning, 3, 64
 Update direction, 163

V

Validation, 91–93
 Value similarity (VS), 572–573
 Variable duration HMM, 829
 Variable elimination, 781
 Variance, 15–17
 Variational approximation methods, 804–805
 Bayesian learning, 640–645
 block methods, 809–813
 Boltzmann machine, 807–809
 multiple-cause networks, 805–807
 noisy-OR model, 805–807
 Variational Bayesian approach
 to Gaussian mixture modeling, 651–654
 to linear regression, 645–651
 Variational bound approximation method, 666–671
 Variational bound Bayesian path, 671–675
 Variational inference techniques, 736–737
 Variational message passing, 812
 Variational method, 670
 VC-dimension of classifier, 91
 Vector space model (VSM), 570–571
 Vector spaces, 109, 394
 Vertex component analysis (VCA) algorithm, 697
 Visual tracking, 867–868

Viterbi algorithm, 825
 Viterbi reestimation, 827–828
 insufficient training data set, 828
 scaling, 828
 Volterra model, 511–514
 Volterra series expansion, 511
v-SVM, 547

W

Weak convergence, 342, 397
 Weierstrass theorem, 510–511
 Welch bound, 424–425
 Well-posed problems, 74
 White Gaussian noise, 238
 White noise
 LS estimator, 237–238
 sequence, 38, 42^f
 Widely-linear APA, 195–196
 Widely linear complex-valued estimation, 113–116
 Widely-linear LMS, 195
 Wide-sense stationary (WSS), 30, 31
 cross-correlation, 34

 real random processes, 119
 two-dimensional random process, 121
 Wiener-Hammerstein model, 512–514, 512^f
 Wiener-Hopf equations, 110
 Wiener model, 511–514, 512^f
 Wireless sensor networks (WSNs), 208
 Wirtinger calculus, 116–118, 175, 1016–1017
 Wirtinger derivative, 117
 Wishart distribution, 601
 Within-class scatter matrix, 296
 Wolfe dual representation, 1029
 Woodbury's matrix inversion formula, 246–247
 WSS. *See* Wide-sense stationary (WSS)

X

X-ray mammography, 2

Y

Yule-Walker equations, 39–40

Z

Zero mean values, 32–33, 106–107