Index

0-1 loss, **104**, 276

Absolute value rectification, 192
Accuracy, 426
Activation function, 170
Active constraint, 95
AdaGrad, 307
ADALINE, *see* adaptive linear element
Adam, 308, 428
Adaptive linear element, 15, 24, 27
Adversarial example, 267
Adversarial training, 268, 270, 533
Affine, 110
AIS, *see* annealed importance sampling
Almost everywhere, 71
Almost sure convergence, 130
Ancestral sampling, 583, 598
ANN, *see* Artificial neural network
Annealed importance sampling, 628, 670, 719
Approximate Bayesian computation, 718
Approximate inference, 586
Artificial intelligence, 1
Artificial neural network, *see* Neural network
ASR, *see* automatic speech recognition
Asymptotically unbiased, 124
Audio, 102, 361, 461
Autoencoder, 4, 357, **505**
Automatic speech recognition, 461

Back-propagation, 203
Back-propagation through time, **385**
Backprop, *see* back-propagation
Bag of words, 474
Bagging, 255
Batch normalization, 266, 428
Bayes error, **117**
Bayes' rule, 70
Bayesian hyperparameter optimization, 439
Bayesian network, *see* directed graphical model
Bayesian probability, 55
Bayesian statistics, **135**
Belief network, *see* directed graphical model
Bernoulli distribution, 62
BFGS, 316
Bias, 124, 229
Bias parameter, 110
Biased importance sampling, 596
Bigram, 465
Binary relation, 485
Block Gibbs sampling, 602
Boltzmann distribution, 573
Boltzmann machine, 573, 656
BPTT, *see* back-propagation through time
Broadcasting, 34
Burn-in, 600

CAE, *see* contractive autoencoder
Calculus of variations, 179
Categorical distribution, *see* multinoulli distribution
CD, *see* contrastive divergence
Centering trick (DBM), 675
Central limit theorem, 63
Chain rule (calculus), 206
Chain rule of probability, 59

Chess, 2
Chord, 582
Chordal graph, 582
Class-based language models, 466
Classical dynamical system, 376
Classification, 100
Clique potential, *see* factor (graphical model)
CNN, *see* convolutional neural network
Collaborative Filtering, 481
Collider, *see* explaining away
Color images, 361
Complex cell, 366
Computational graph, 204
Computer vision, 455
Concept drift, 541
Condition number, 279
Conditional computation, *see* dynamic structure
Conditional independence, xiii, 60
Conditional probability, 59
Conditional RBM, 687
Connectionism, 17, 446
Connectionist temporal classification, 463
Consistency, 130, 516
Constrained optimization, 93, 237
Content-based addressing, 422
Content-based recommender systems, 483
Context-specific independence, 576
Contextual bandits, 483
Continuation methods, 328
Contractive autoencoder, 524
Contrast, 457
Contrastive divergence, 291, 613, 674
Convex optimization, 141
Convolution, 331, 685
Convolutional network, 16
Convolutional neural network, 252, **331**, 428, 463
Coordinate descent, 322, 673
Correlation, 61
Cost function, *see* objective function
Covariance, xiii, 61
Covariance matrix, 62
Coverage, 427
Critical temperature, 606
Cross-correlation, 333
Cross-entropy, **75**, 132
Cross-validation, 122
CTC, *see* connectionist temporal classification
Curriculum learning, 329
Curse of dimensionality, 154
Cyc, 2

D-separation, 575
DAE, *see* denoising autoencoder
Data generating distribution, **111**, 131
Data generating process, 111
Data parallelism, 450
Dataset, 105
Dataset augmentation, 270, 460
DBM, *see* deep Boltzmann machine
DCGAN, 554, 555, 703
Decision tree, **145**, 551
Decoder, 4
Deep belief network, 27, 532, 634, 659, 662, 686, 694
Deep Blue, 2
Deep Boltzmann machine, 24, 27, 532, 634, 655, 659, 665, 674, 686
Deep feedforward network, 167, 428
Deep learning, 2, 5
Denoising autoencoder, 513, 691
Denoising score matching, 622
Density estimation, 103
Derivative, xiii, 83
Design matrix, **106**
Detector layer, 340
Determinant, xii
Diagonal matrix, 41
Differential entropy, 74, 649
Dirac delta function, 65
Directed graphical model, 77, 510, 566, 694
Directional derivative, 85
Discriminative fine-tuning, *see* supervised fine-tuning
Discriminative RBM, 688
Distributed representation, 17, 150, 549
Domain adaptation, 539

Dot product, 34, 141
Double backprop, 270
Doubly block circulant matrix, 334
Dream sleep, 612, 655
DropConnect, 265
Dropout, **257**, 428, 433, 434, 674, 691
Dynamic structure, 451, 452

E-step, 637
Early stopping, 246, 249, 272, 273, 428
EBM, *see* energy-based model
Echo state network, 24, 27, 406
Effective capacity, 114
Eigendecomposition, 42
Eigenvalue, 42
Eigenvector, 42
ELBO, *see* evidence lower bound
Element-wise product, *see* Hadamard product, *see* Hadamard product
EM, *see* expectation maximization
Embedding, 519
Empirical distribution, 66
Empirical risk, 276
Empirical risk minimization, 276
Encoder, 4
Energy function, 572
Energy-based model, 572, 598, 656, 665
Ensemble methods, 255
Epoch, 247
Equality constraint, 94
Equivariance, 339
Error function, *see* objective function
ESN, *see* echo state network
Euclidean norm, 39
Euler-Lagrange equation, 649
Evidence lower bound, 636, 663
Example, 99
Expectation, 60
Expectation maximization, 637
Expected value, *see* expectation
Explaining away, 577, 634, 647
Exploitation, 484
Exploration, 484
Exponential distribution, **65**

F-score, 426
Factor (graphical model), 570
Factor analysis, 493
Factor graph, 582
Factors of variation, 4
Feature, 99
Feature selection, 236
Feedforward neural network, 167
Fine-tuning, 324
Finite differences, 442
Forget gate, 306
Forward propagation, 203
Fourier transform, 361, 363
Fovea, 367
FPCD, 617
Free energy, **574**, 682
`Freebase`, 486
Frequentist probability, 55
Frequentist statistics, **135**
Frobenius norm, 46
Fully-visible Bayes network, 707
Functional derivatives, 648
FVBN, *see* fully-visible Bayes network

Gabor function, 369
GANs, *see* generative adversarial networks
Gated recurrent unit, 428
Gaussian distribution, *see* normal distribution
Gaussian kernel, 142
Gaussian mixture, 67, 188
GCN, *see* global contrast normalization
`GeneOntology`, 486
Generalization, 110
Generalized Lagrange function, *see* generalized Lagrangian
Generalized Lagrangian, 94
Generative adversarial networks, 691, 702
Generative moment matching networks, 705
Generator network, 695
Gibbs distribution, 571
Gibbs sampling, 584, 602
Global contrast normalization, 457
GPU, *see* graphics processing unit
Gradient, 84

Gradient clipping, 289, 417
Gradient descent, 83, 85
Graph, xii
Graphical model, *see* structured probabilistic model
Graphics processing unit, 447
Greedy algorithm, 324
Greedy layer-wise unsupervised pretraining, 531
Greedy supervised pretraining, 324
Grid search, 435

Hadamard product, xii, 34
Hard tanh, 196
Harmonium, *see* restricted Boltzmann machine
Harmony theory, 574
Helmholtz free energy, *see* evidence lower bound
Hessian, 223
Hessian matrix, xiii, 87
Heteroscedastic, 187
Hidden layer, 6, 167
Hill climbing, 86
Hyperparameter optimization, 435
Hyperparameters, 120, 433
Hypothesis space, 112, 118

i.i.d. assumptions, 111, 122, 267
Identity matrix, 36
ILSVRC, *see* ImageNet Large-Scale Visual Recognition Challenge
ImageNet Large-Scale Visual Recognition Challenge, 23
Immorality, 580
Importance sampling, 595, 627, 700
Importance weighted autoencoder, 700
Independence, xiii, 60
Independent and identically distributed, *see* i.i.d. assumptions
Independent component analysis, 494
Independent subspace analysis, 496
Inequality constraint, 94
Inference, 565, 586, 634, 636, 638, 641, 651, 653
Information retrieval, 528
Initialization, 301
Integral, xiii
Invariance, 343
Isotropic, 65

Jacobian matrix, xiii, 72, 86
Joint probability, 57

k-means, 365, 549
k-nearest neighbors, **143**, 551
Karush-Kuhn-Tucker conditions, 95, 237
Karush–Kuhn–Tucker, 94
Kernel (convolution), 332, 333
Kernel machine, 551
Kernel trick, 141
KKT, *see* Karush–Kuhn–Tucker
KKT conditions, *see* Karush-Kuhn-Tucker conditions
KL divergence, *see* Kullback-Leibler divergence
Knowledge base, 2, 486
Krylov methods, 224
Kullback-Leibler divergence, xiii, **74**

Label smoothing, 243
Lagrange multipliers, 94, 649
Lagrangian, *see* generalized Lagrangian
LAPGAN, 704
Laplace distribution, **65**, 499, 500
Latent variable, 67
Layer (neural network), 167
LCN, *see* local contrast normalization
Leaky ReLU, 192
Leaky units, 409
Learning rate, 85
Line search, 85, 86, 93
Linear combination, 37
Linear dependence, 38
Linear factor models, 492
Linear regression, **107**, 110, 140
Link prediction, 487
Lipschitz constant, 92
Lipschitz continuous, 92
Liquid state machine, 406

Local conditional probability distribution, 567
Local contrast normalization, 459
Logistic regression, 3, **140**, 140
Logistic sigmoid, 7, 67
Long short-term memory, 18, 25, 306, **411**, 428
Loop, 582
Loopy belief propagation, 588
Loss function, *see* objective function
L norm, 39
LSTM, *see* long short-term memory

M-step, 637
Machine learning, 2
Machine translation, 101
Main diagonal, 33
Manifold, 160
Manifold hypothesis, 161
Manifold learning, 161
Manifold tangent classifier, 270
MAP approximation, 138, 508
Marginal probability, 58
Markov chain, 598
Markov chain Monte Carlo, 598
Markov network, *see* undirected model
Markov random field, *see* undirected model
Matrix, xi, xii, 32
Matrix inverse, 36
Matrix product, 34
Max norm, 40
Max pooling, 340
Maximum likelihood, **131**
Maxout, 192, 428
MCMC, *see* Markov chain Monte Carlo
Mean field, 641, 642, 674
Mean squared error, 108
Measure theory, 71
Measure zero, 71
Memory network, 419, 421
Method of steepest descent, *see* gradient descent
Minibatch, 279
Missing inputs, 100
Mixing (Markov chain), 604
Mixture density networks, 188
Mixture distribution, 66
Mixture model, 188, 513
Mixture of experts, 453, 551
MLP, *see* multilayer perception
MNIST, 21, 22, 674
Model averaging, 255
Model compression, 451
Model identifiability, 284
Model parallelism, 450
Moment matching, 705
Moore-Penrose pseudoinverse, 45, 240
Moralized graph, 580
MP-DBM, *see* multi-prediction DBM
MRF (Markov Random Field), *see* undirected model
MSE, *see* mean squared error
Multi-modal learning, 542
Multi-prediction DBM, 676
Multi-task learning, 245, 541
Multilayer perception, 5
Multilayer perceptron, 27
Multinomial distribution, 62
Multinoulli distribution, 62

n-gram, **464**
NADE, 710
Naive Bayes, 3
Nat, 73
Natural image, 562
Natural language processing, 464
Nearest neighbor regression, **115**
Negative definite, 89
Negative phase, 473, 609, 611
Neocognitron, 16, 24, 27, 368
Nesterov momentum, 300
Netflix Grand Prize, 256, 482
Neural language model, 466, 479
Neural network, 13
Neural Turing machine, 421
Neuroscience, 15
Newton's method, 89, 310
NLM, *see* neural language model
NLP, *see* natural language processing
No free lunch theorem, 116

Noise-contrastive estimation, 623
Non-parametric model, **114**
Norm, xiv, 39
Normal distribution, 63, 64, 125
Normal equations, **109**, 109, 112, 234
Normalized initialization, 303
Numerical differentiation, *see* finite differences

Object detection, 456
Object recognition, 456
Objective function, 82
OMP-k, *see* orthogonal matching pursuit
One-shot learning, 541
Operation, 204
Optimization, 80, 82
Orthodox statistics, *see* frequentist statistics
Orthogonal matching pursuit, 27, **254**
Orthogonal matrix, 42
Orthogonality, 41
Output layer, 167

Parallel distributed processing, 17
Parameter initialization, 301, 408
Parameter sharing, 251, 336, 374, 376, 389
Parameter tying, *see* Parameter sharing
Parametric model, **114**
Parametric ReLU, 192
Partial derivative, 84
Partition function, 571, 608, 671
PCA, *see* principal components analysis
PCD, *see* stochastic maximum likelihood
Perceptron, 15, 27
Persistent contrastive divergence, *see* stochastic maximum likelihood
Perturbation analysis, *see* reparametrization trick
Point estimator, 122
Policy, 483
Pooling, 331, 685
Positive definite, 89
Positive phase, 473, 609, 611, 658, 670
Precision, 426
Precision (of a normal distribution), 63, 65
Predictive sparse decomposition, 526
Preprocessing, 456
Pretraining, 324, 531
Primary visual cortex, 366
Principal components analysis, 48, 146–148, 493, 634
Prior probability distribution, **135**
Probabilistic max pooling, 685
Probabilistic PCA, 493, 494, 635
Probability density function, 58
Probability distribution, 56
Probability mass function, 56
Probability mass function estimation, 103
Product of experts, 573
Product rule of probability, *see* chain rule of probability
PSD, *see* predictive sparse decomposition
Pseudolikelihood, 618

Quadrature pair, 370
Quasi-Newton condition, 316
Quasi-Newton methods, 316

Radial basis function, 196
Random search, 437
Random variable, 56
Ratio matching, 621
RBF, 196
RBM, *see* restricted Boltzmann machine
Recall, 426
Receptive field, 338
Recommender Systems, 481
Rectified linear unit, 171, 192, 428, 510
Recurrent network, 27
Recurrent neural network, 379
Regression, 101
Regularization, **120**, 120, 177, 228, 433
Regularizer, 119
REINFORCE, 691
Reinforcement learning, 25, 106, 483, 691
Relational database, 486
Reparametrization trick, 690
Representation learning, 3
Representational capacity, 114
Restricted Boltzmann machine, 357, 462, 482, 590, 634, 658, 659, 674, 678,

680, 683, 685
Ridge regression, *see* weight decay
Risk, 275
RNN-RBM, 688

Saddle points, 285
Sample mean, 125
Scalar, xi, xii, 31
Score matching, 516, 620
Secant condition, 316
Second derivative, 86
Second derivative test, 89
Self-information, 73
Semantic hashing, 528
Semi-supervised learning, 244
Separable convolution, 363
Separation (probabilistic modeling), 575
Set, xii
SGD, *see* stochastic gradient descent
Shannon entropy, xiii, 73
Shortlist, 469
Sigmoid, xiv, *see* logistic sigmoid
Sigmoid belief network, 27
Simple cell, 366
Singular value, *see* singular value decomposition
Singular value decomposition, 44, 148, 482
Singular vector, *see* singular value decomposition
Slow feature analysis, 496
SML, *see* stochastic maximum likelihood
Softmax, 183, 421, 453
Softplus, xiv, 68, 196
Spam detection, 3
Sparse coding, 322, 357, 499, 634, 694
Sparse initialization, 304, 408
Sparse representation, 146, 226, 253, 508, 559
Spearmint, 439
Spectral radius, 407
Speech recognition, *see* automatic speech recognition
Sphering, *see* whitening
Spike and slab restricted Boltzmann machine, 683
SPN, *see* sum-product network
Square matrix, 38
ssRBM, *see* spike and slab restricted Boltzmann machine
Standard deviation, 61
Standard error, 127
Standard error of the mean, 128, 278
Statistic, 122
Statistical learning theory, 110
Steepest descent, *see* gradient descent
Stochastic back-propagation, *see* reparametrization trick
Stochastic gradient descent, 15, 150, 279, **294**, 674
Stochastic maximum likelihood, 615, 674
Stochastic pooling, 265
Structure learning, 585
Structured output, 101, 687
Structured probabilistic model, 77, 561
Sum rule of probability, 58
Sum-product network, 556
Supervised fine-tuning, 532, 664
Supervised learning, **105**
Support vector machine, 140
Surrogate loss function, 276
SVD, *see* singular value decomposition
Symmetric matrix, 41, 43

Tangent distance, 269
Tangent plane, 519
Tangent prop, 269
TDNN, *see* time-delay neural network
Teacher forcing, 383, 384
Tempering, 606
Template matching, 141
Tensor, xi, xii, 33
Test set, 110
Tikhonov regularization, *see* weight decay
Tiled convolution, 353
Time-delay neural network, 369, 375
Toeplitz matrix, 334
Topographic ICA, 496
Trace operator, 46
Training error, 110
Transcription, 101

Transfer learning, 539
Transpose, xii, 33
Triangle inequality, 39
Triangulated graph, *see* chordal graph
Trigram, 465

Unbiased, 124
Undirected graphical model, 77, 510
Undirected model, 569
Uniform distribution, 57
Unigram, 465
Unit norm, 41
Unit vector, 41
Universal approximation theorem, 197
Universal approximator, 556
Unnormalized probability distribution, 570
Unsupervised learning, **105**, 146
Unsupervised pretraining, 462, 531

V-structure, *see* explaining away
V1, 366
VAE, *see* variational autoencoder
Vapnik-Chervonenkis dimension, 114
Variance, xiii, 61, 229
Variational autoencoder, 691, **698**
Variational derivatives, *see* functional derivatives
Variational free energy, *see* evidence lower bound
VC dimension, *see* Vapnik-Chervonenkis dimension
Vector, xi, xii, 32
Virtual adversarial examples, 268
Visible layer, 6
Volumetric data, 361

Wake-sleep, 654, 663
Weight decay, **118**, 177, 231, 434
Weight space symmetry, 284
Weights, 15, 107
Whitening, 458
Wikibase, 486
`Wikibase`, 486
Word embedding, 467
Word-sense disambiguation, 487
`WordNet`, 486

Zero-data learning, *see* zero-shot learning
Zero-shot learning, 541