Index

0-1 loss, **102**, 274

Absolute value rectification, 191
Accuracy, 420
Activation function, 169
Active constraint, 94
AdaGrad, 305
ADALINE, *see* adaptive linear element
Adam, 307, 422
Adaptive linear element, 15, 23, 26
Adversarial example, 265
Adversarial training, 266, 268, 526
Affine, 109
AIS, *see* annealed importance sampling
Almost everywhere, 70
Almost sure convergence, 128
Ancestral sampling, 576, 591
ANN, *see* Artificial neural network
Annealed importance sampling, 621, 662, 711
Approximate Bayesian computation, 710
Approximate inference, 579
Artificial intelligence, 1
Artificial neural network, *see* Neural network
ASR, *see* automatic speech recognition
Asymptotically unbiased, 123
Audio, 101, 357, 455
Autoencoder, 4, 353, **498**
Automatic speech recognition, 455

Back-propagation, 201
Back-propagation through time, **381**
Backprop, *see* back-propagation
Bag of words, 467
Bagging, 252
Batch normalization, 264, 422
Bayes error, **116**
Bayes' rule, 69
Bayesian hyperparameter optimization, 433
Bayesian network, *see* directed graphical model
Bayesian probability, 54
Bayesian statistics, **134**
Belief network, *see* directed graphical model
Bernoulli distribution, 61
BFGS, 314
Bias, 123, 227
Bias parameter, 109
Biased importance sampling, 589
Bigram, 458
Binary relation, 478
Block Gibbs sampling, 595
Boltzmann distribution, 566
Boltzmann machine, 566, 648
BPTT, *see* back-propagation through time
Broadcasting, 33
Burn-in, 593

CAE, *see* contractive autoencoder
Calculus of variations, 178
Categorical distribution, *see* multinoulli distribution
CD, *see* contrastive divergence
Centering trick (DBM), 667
Central limit theorem, 63
Chain rule (calculus), 203
Chain rule of probability, 58

Chess, 2
Chord, 575
Chordal graph, 575
Class-based language models, 460
Classical dynamical system, 372
Classification, 99
Clique potential, *see* factor (graphical model)
CNN, *see* convolutional neural network
Collaborative Filtering, 474
Collider, *see* explaining away
Color images, 357
Complex cell, 362
Computational graph, 202
Computer vision, 449
Concept drift, 533
Condition number, 277
Conditional computation, *see* dynamic structure
Conditional independence, xiii, 59
Conditional probability, 58
Conditional RBM, 679
Connectionism, 17, 440
Connectionist temporal classification, 457
Consistency, 128, 509
Constrained optimization, 92, 235
Content-based addressing, 416
Content-based recommender systems, 475
Context-specific independence, 569
Contextual bandits, 476
Continuation methods, 324
Contractive autoencoder, 516
Contrast, 451
Contrastive divergence, 289, 606, 666
Convex optimization, 140
Convolution, 327, 677
Convolutional network, 16
Convolutional neural network, 250, **327**, 422, 456
Coordinate descent, 319, 665
Correlation, 60
Cost function, *see* objective function
Covariance, xiii, 60
Covariance matrix, 61
Coverage, 421
Critical temperature, 599
Cross-correlation, 329
Cross-entropy, **74**, 131
Cross-validation, 121
CTC, *see* connectionist temporal classification
Curriculum learning, 326
Curse of dimensionality, 153
Cyc, 2

D-separation, 568
DAE, *see* denoising autoencoder
Data generating distribution, **110**, 130
Data generating process, 110
Data parallelism, 444
Dataset, 103
Dataset augmentation, 268, 454
DBM, *see* deep Boltzmann machine
DCGAN, 547, 548, 695
Decision tree, **144**, 544
Decoder, 4
Deep belief network, 26, 525, 626, 651, 654, 678, 686
Deep Blue, 2
Deep Boltzmann machine, 23, 26, 525, 626, 647, 651, 657, 666, 678
Deep feedforward network, 166, 422
Deep learning, 2, 5
Denoising autoencoder, 506, 683
Denoising score matching, 615
Density estimation, 102
Derivative, xiii, 82
Design matrix, **105**
Detector layer, 336
Determinant, xii
Diagonal matrix, 40
Differential entropy, 73, 641
Dirac delta function, 64
Directed graphical model, 76, 503, 559, 685
Directional derivative, 84
Discriminative fine-tuning, *see* supervised fine-tuning
Discriminative RBM, 680
Distributed representation, 17, 149, 542
Domain adaptation, 532

Dot product, 33, 139
Double backprop, 268
Doubly block circulant matrix, 330
Dream sleep, 605, 647
DropConnect, 263
Dropout, **255**, 422, 427, 428, 666, 683
Dynamic structure, 445

E-step, 629
Early stopping, 244, 246, 270, 271, 422
EBM, *see* energy-based model
Echo state network, 23, 26, 401
Effective capacity, 113
Eigendecomposition, 41
Eigenvalue, 41
Eigenvector, 41
ELBO, *see* evidence lower bound
Element-wise product, *see* Hadamard product, *see* Hadamard product
EM, *see* expectation maximization
Embedding, 512
Empirical distribution, 65
Empirical risk, 274
Empirical risk minimization, 274
Encoder, 4
Energy function, 565
Energy-based model, 565, 591, 648, 657
Ensemble methods, 252
Epoch, 244
Equality constraint, 93
Equivariance, 335
Error function, *see* objective function
ESN, *see* echo state network
Euclidean norm, 38
Euler-Lagrange equation, 641
Evidence lower bound, 628, 655
Example, 98
Expectation, 59
Expectation maximization, 629
Expected value, *see* expectation
Explaining away, 570, 626, 639
Exploitation, 477
Exploration, 477
Exponential distribution, **64**

F-score, 420
Factor (graphical model), 563
Factor analysis, 486
Factor graph, 575
Factors of variation, 4
Feature, 98
Feature selection, 234
Feedforward neural network, 166
Fine-tuning, 321
Finite differences, 436
Forget gate, 304
Forward propagation, 201
Fourier transform, 357, 359
Fovea, 363
FPCD, 610
Free energy, **567**, 674
`Freebase`, 479
Frequentist probability, 54
Frequentist statistics, **134**
Frobenius norm, 45
Fully-visible Bayes network, 699
Functional derivatives, 640
FVBN, *see* fully-visible Bayes network

Gabor function, 365
GANs, *see* generative adversarial networks
Gated recurrent unit, 422
Gaussian distribution, *see* normal distribution
Gaussian kernel, 140
Gaussian mixture, 66, 187
GCN, *see* global contrast normalization
`GeneOntology`, 479
Generalization, 109
Generalized Lagrange function, *see* generalized Lagrangian
Generalized Lagrangian, 93
Generative adversarial networks, 683, 693
Generative moment matching networks, 696
Generator network, 687
Gibbs distribution, 564
Gibbs sampling, 577, 595
Global contrast normalization, 451
GPU, *see* graphics processing unit
Gradient, 83

Gradient clipping, 287, 411
Gradient descent, 82, 84
Graph, xii
Graphical model, *see* structured probabilistic model
Graphics processing unit, 441
Greedy algorithm, 321
Greedy layer-wise unsupervised pretraining, 524
Greedy supervised pretraining, 321
Grid search, 429

Hadamard product, xii, 33
Hard tanh, 195
Harmonium, *see* restricted Boltzmann machine
Harmony theory, 567
Helmholtz free energy, *see* evidence lower bound
Hessian, 221
Hessian matrix, xiii, 86
Heteroscedastic, 186
Hidden layer, 6, 166
Hill climbing, 85
Hyperparameter optimization, 429
Hyperparameters, 119, 427
Hypothesis space, 111, 117

i.i.d. assumptions, 110, 121, 265
Identity matrix, 35
ILSVRC, *see* ImageNet Large Scale Visual Recognition Challenge
ImageNet Large Scale Visual Recognition Challenge, 22
Immorality, 573
Importance sampling, 588, 620, 691
Importance weighted autoencoder, 691
Independence, xiii, 59
Independent and identically distributed, *see* i.i.d. assumptions
Independent component analysis, 487
Independent subspace analysis, 489
Inequality constraint, 93
Inference, 558, 579, 626, 628, 630, 633, 643, 646
Information retrieval, 520
Initialization, 298
Integral, xiii
Invariance, 339
Isotropic, 64

Jacobian matrix, xiii, 71, 85
Joint probability, 56

k-means, 361, 542
k-nearest neighbors, **141**, 544
Karush-Kuhn-Tucker conditions, 94, 235
Karush–Kuhn–Tucker, 93
Kernel (convolution), 328, 329
Kernel machine, 544
Kernel trick, 139
KKT, *see* Karush–Kuhn–Tucker
KKT conditions, *see* Karush-Kuhn-Tucker conditions
KL divergence, *see* Kullback-Leibler divergence
Knowledge base, 2, 479
Krylov methods, 222
Kullback-Leibler divergence, xiii, **73**

Label smoothing, 241
Lagrange multipliers, 93, 641
Lagrangian, *see* generalized Lagrangian
LAPGAN, 695
Laplace distribution, **64**, 492
Latent variable, 66
Layer (neural network), 166
LCN, *see* local contrast normalization
Leaky ReLU, 191
Leaky units, 404
Learning rate, 84
Line search, 84, 85, 92
Linear combination, 36
Linear dependence, 37
Linear factor models, 485
Linear regression, **106**, 109, 138
Link prediction, 480
Lipschitz constant, 91
Lipschitz continuous, 91
Liquid state machine, 401

Local conditional probability distribution, 560
Local contrast normalization, 452
Logistic regression, 3, **138**, 139
Logistic sigmoid, 7, 66
Long short-term memory, 18, 24, 304, **407**, 422
Loop, 575
Loopy belief propagation, 581
Loss function, *see* objective function
L^p norm, 38
LSTM, *see* long short-term memory

M-step, 629
Machine learning, 2
Machine translation, 100
Main diagonal, 32
Manifold, 159
Manifold hypothesis, 160
Manifold learning, 160
Manifold tangent classifier, 268
MAP approximation, 137, 501
Marginal probability, 57
Markov chain, 591
Markov chain Monte Carlo, 591
Markov network, *see* undirected model
Markov random field, *see* undirected model
Matrix, xi, xii, 31
Matrix inverse, 35
Matrix product, 33
Max norm, 39
Max pooling, 336
Maximum likelihood, **130**
Maxout, 191, 422
MCMC, *see* Markov chain Monte Carlo
Mean field, 633, 634, 666
Mean squared error, 107
Measure theory, 70
Measure zero, 70
Memory network, 413, 415
Method of steepest descent, *see* gradient descent
Minibatch, 277
Missing inputs, 99
Mixing (Markov chain), 597
Mixture density networks, 187
Mixture distribution, 65
Mixture model, 187, 506
Mixture of experts, 446, 544
MLP, *see* multilayer perception
MNIST, 20, 21, 666
Model averaging, 252
Model compression, 444
Model identifiability, 282
Model parallelism, 444
Moment matching, 696
Moore-Penrose pseudoinverse, 44, 237
Moralized graph, 573
MP-DBM, *see* multi-prediction DBM
MRF (Markov Random Field), *see* undirected model
MSE, *see* mean squared error
Multi-modal learning, 535
Multi-prediction DBM, 668
Multi-task learning, 242, 533
Multilayer perception, 5
Multilayer perceptron, 26
Multinomial distribution, 61
Multinoulli distribution, 61

n-gram, **458**
NADE, 702
Naive Bayes, 3
Nat, 72
Natural image, 555
Natural language processing, 457
Nearest neighbor regression, **114**
Negative definite, 88
Negative phase, 466, 602, 604
Neocognitron, 16, 23, 26, 364
Nesterov momentum, 298
Netflix Grand Prize, 255, 475
Neural language model, 460, 472
Neural network, 13
Neural Turing machine, 415
Neuroscience, 15
Newton's method, 88, 309
NLM, *see* neural language model
NLP, *see* natural language processing
No free lunch theorem, 115

Noise-contrastive estimation, 616
Non-parametric model, **113**
Norm, xiv, 38
Normal distribution, 62, 63, 124
Normal equations, **108**, 108, 111, 232
Normalized initialization, 301
Numerical differentiation, *see* finite differences

Object detection, 449
Object recognition, 449
Objective function, 81
OMP-k, *see* orthogonal matching pursuit
One-shot learning, 534
Operation, 202
Optimization, 79, 81
Orthodox statistics, *see* frequentist statistics
Orthogonal matching pursuit, 26, **252**
Orthogonal matrix, 41
Orthogonality, 40
Output layer, 166

Parallel distributed processing, 17
Parameter initialization, 298, 403
Parameter sharing, 249, 332, 370, 372, 386
Parameter tying, *see* Parameter sharing
Parametric model, **113**
Parametric ReLU, 191
Partial derivative, 83
Partition function, 564, 601, 663
PCA, *see* principal components analysis
PCD, *see* stochastic maximum likelihood
Perceptron, 15, 26
Persistent contrastive divergence, *see* stochastic maximum likelihood
Perturbation analysis, *see* reparametrization trick
Point estimator, 121
Policy, 476
Pooling, 327, 677
Positive definite, 88
Positive phase, 466, 602, 604, 650, 662
Precision, 420
Precision (of a normal distribution), 62, 64
Predictive sparse decomposition, 519
Preprocessing, 450
Pretraining, 320, 524
Primary visual cortex, 362
Principal components analysis, 47, 145, 146, 486, 626
Prior probability distribution, **134**
Probabilistic max pooling, 677
Probabilistic PCA, 486, 487, 627
Probability density function, 57
Probability distribution, 55
Probability mass function, 55
Probability mass function estimation, 102
Product of experts, 566
Product rule of probability, *see* chain rule of probability
PSD, *see* predictive sparse decomposition
Pseudolikelihood, 611

Quadrature pair, 366
Quasi-Newton methods, 314

Radial basis function, 195
Random search, 431
Random variable, 55
Ratio matching, 614
RBF, 195
RBM, *see* restricted Boltzmann machine
Recall, 420
Receptive field, 334
Recommender Systems, 474
Rectified linear unit, 170, 191, 422, 503
Recurrent network, 26
Recurrent neural network, 375
Regression, 99
Regularization, **119**, 119, 176, 226, 427
Regularizer, 118
REINFORCE, 683
Reinforcement learning, 24, 105, 476, 683
Relational database, 479
Relations, 478
Reparametrization trick, 682
Representation learning, 3
Representational capacity, 113
Restricted Boltzmann machine, 353, 456, 475, 583, 626, 650, 651, 666, 670,

672, 674, 677
Ridge regression, *see* weight decay
Risk, 273
RNN-RBM, 679

Saddle points, 283
Sample mean, 124
Scalar, xi, xii, 30
Score matching, 509, 613
Second derivative, 85
Second derivative test, 88
Self-information, 72
Semantic hashing, 521
Semi-supervised learning, 241
Separable convolution, 359
Separation (probabilistic modeling), 568
Set, xii
SGD, *see* stochastic gradient descent
Shannon entropy, xiii, 73
Shortlist, 462
Sigmoid, xiv, *see* logistic sigmoid
Sigmoid belief network, 26
Simple cell, 362
Singular value, *see* singular value decomposition
Singular value decomposition, 43, 146, 475
Singular vector, *see* singular value decomposition
Slow feature analysis, 489
SML, *see* stochastic maximum likelihood
Softmax, 182, 415, 446
Softplus, xiv, 67, 195
Spam detection, 3
Sparse coding, 319, 353, 492, 626, 686
Sparse initialization, 302, 403
Sparse representation, 145, 224, 251, 501, 552
Spearmint, 433
Spectral radius, 401
Speech recognition, *see* automatic speech recognition
Sphering, *see* whitening
Spike and slab restricted Boltzmann machine, 674
SPN, *see* sum-product network
Square matrix, 37
ssRBM, *see* spike and slab restricted Boltzmann machine
Standard deviation, 60
Standard error, 126
Standard error of the mean, 126, 276
Statistic, 121
Statistical learning theory, 109
Steepest descent, *see* gradient descent
Stochastic back-propagation, *see* reparametrization trick
Stochastic gradient descent, 15, 149, 277, **292**, 666
Stochastic maximum likelihood, 608, 666
Stochastic pooling, 263
Structure learning, 578
Structured output, 100, 679
Structured probabilistic model, 76, 554
Sum rule of probability, 57
Sum-product network, 549
Supervised fine-tuning, 525, 656
Supervised learning, **104**
Support vector machine, 139
Surrogate loss function, 274
SVD, *see* singular value decomposition
Symmetric matrix, 40, 42

Tangent distance, 267
Tangent plane, 511
Tangent prop, 267
TDNN, *see* time-delay neural network
Teacher forcing, 379, 380
Tempering, 599
Template matching, 140
Tensor, xi, xii, 32
Test set, 109
Tikhonov regularization, *see* weight decay
Tiled convolution, 349
Time-delay neural network, 364, 371
Toeplitz matrix, 330
Topographic ICA, 489
Trace operator, 45
Training error, 109
Transcription, 100
Transfer learning, 532

Transpose, xii, 32
Triangle inequality, 38
Triangulated graph, *see* chordal graph
Trigram, 458

Unbiased, 123
Undirected graphical model, 76, 503
Undirected model, 562
Uniform distribution, 56
Unigram, 458
Unit norm, 40
Unit vector, 40
Universal approximation theorem, 196
Universal approximator, 549
Unnormalized probability distribution, 563
Unsupervised learning, **104**, 144
Unsupervised pretraining, 456, 524

V-structure, *see* explaining away
V1, 362
VAE, *see* variational autoencoder
Vapnik-Chervonenkis dimension, 113
Variance, xiii, 60, 227
Variational autoencoder, 683, **690**
Variational derivatives, *see* functional derivatives
Variational free energy, *see* evidence lower bound
VC dimension, *see* Vapnik-Chervonenkis dimension
Vector, xi, xii, 31
Virtual adversarial examples, 266
Visible layer, 6
Volumetric data, 357

Wake-sleep, 646, 655
Weight decay, **117**, 176, 229, 428
Weight space symmetry, 282
Weights, 15, 106
Whitening, 452
Wikibase, 479
`Wikibase`, 479
Word embedding, 460
Word-sense disambiguation, 480
`WordNet`, 479

Zero-data learning, *see* zero-shot learning
Zero-shot learning, 534