

Index

- 0-1 loss, [100](#), [270](#)
- Absolute value rectification, [187](#)
- Accuracy, [414](#)
- Activation function, [165](#)
- Active constraint, [92](#)
- AdaGrad, [300](#)
- ADALINE, *see* adaptive linear element
- Adam, [302](#), [416](#)
- Adaptive linear element, [14](#), [21](#), [23](#)
- Adversarial example, [263](#)
- Adversarial training, [263](#), [267](#), [522](#)
- Affine, [107](#)
- AIS, *see* annealed importance sampling
- Almost everywhere, [68](#)
- Almost sure convergence, [126](#)
- Ancestral sampling, [572](#), [587](#)
- ANN, *see* Artificial neural network
- Annealed importance sampling, [617](#), [658](#), [705](#)
- Approximate Bayesian computation, [705](#)
- Approximate inference, [575](#)
- Artificial intelligence, [1](#)
- Artificial neural network, *see* Neural network
- ASR, *see* automatic speech recognition
- Asymptotically unbiased, [121](#)
- Audio, [99](#), [350](#), [449](#)
- Autoencoder, [4](#), [347](#), [493](#)
- Automatic speech recognition, [449](#)
- Back-propagation, [198](#)
- Back-propagation through time, [375](#)
- Backprop, *see* back-propagation
- Bag of words, [461](#)
- Bagging, [250](#)
- Batch normalization, [262](#), [416](#)
- Bayes error, [114](#)
- Bayes' rule, [67](#)
- Bayesian hyperparameter optimization, [426](#)
- Bayesian network, *see* directed graphical model
- Bayesian probability, [52](#)
- Bayesian statistics, [132](#)
- Belief network, *see* directed graphical model
- Bernoulli distribution, [59](#)
- BFGS, [309](#)
- Bias, [121](#), [223](#)
- Bias parameter, [107](#)
- Biased importance sampling, [585](#)
- Bigram, [452](#)
- Binary relation, [473](#)
- Block Gibbs sampling, [591](#)
- Boltzmann distribution, [562](#)
- Boltzmann machine, [562](#), [644](#)
- BPTT, *see* back-propagation through time
- Broadcasting, [31](#)
- Burn-in, [589](#)
- CAE, *see* contractive autoencoder
- Calculus of variations, [173](#)
- Categorical distribution, *see* multinoulli distribution
- CD, *see* contrastive divergence
- Centering trick (DBM), [663](#)
- Central limit theorem, [61](#)
- Chain rule (calculus), [199](#)
- Chain rule of probability, [56](#)

- Chess, [2](#)
- Chord, [569](#)
- Chordal graph, [569](#)
- Class-based language models, [454](#)
- Classical dynamical system, [366](#)
- Classification, [97](#)
- Clique potential, *see* factor (graphical model)
- CNN, *see* convolutional neural network
- Collaborative Filtering, [469](#)
- Collider, *see* explaining away
- Color images, [350](#)
- Complex cell, [356](#)
- Computational graph, [199](#)
- Computer vision, [443](#)
- Concept drift, [530](#)
- Condition number, [273](#)
- Conditional computation, *see* dynamic structure
- Conditional independence, [xiii](#), [57](#)
- Conditional probability, [56](#)
- Conditional RBM, [675](#)
- Connectionism, [16](#), [434](#)
- Connectionist temporal classification, [451](#)
- Consistency, [126](#), [503](#)
- Constrained optimization, [90](#), [231](#)
- Content-based addressing, [409](#)
- Content-based recommender systems, [470](#)
- Context-specific independence, [565](#)
- Contextual bandits, [470](#)
- Continuation methods, [320](#)
- Contractive autoencoder, [512](#)
- Contrast, [445](#)
- Contrastive divergence, [284](#), [602](#), [661](#)
- Convex optimization, [138](#)
- Convolution, [323](#), [672](#)
- Convolutional network, [15](#)
- Convolutional neural network, [248](#), [323](#), [416](#), [450](#)
- Coordinate descent, [314](#), [659](#)
- Correlation, [58](#)
- Cost function, *see* objective function
- Covariance, [xiii](#), [58](#)
- Covariance matrix, [59](#)
- Coverage, [415](#)
- Critical temperature, [595](#)
- Cross-correlation, [325](#)
- Cross-entropy, [72](#), [129](#)
- Cross-validation, [119](#)
- CTC, *see* connectionist temporal classification
- Curriculum learning, [321](#)
- Curse of dimensionality, [151](#)
- Cyc, [2](#)
- D-separation, [564](#)
- DAE, *see* denoising autoencoder
- Data generating distribution, [108](#), [128](#)
- Data generating process, [108](#)
- Data parallelism, [438](#)
- Dataset, [101](#)
- Dataset augmentation, [266](#), [448](#)
- DBM, *see* deep Boltzmann machine
- DCGAN, [543](#), [544](#), [690](#)
- Decision tree, [140](#), [540](#)
- Decoder, [4](#)
- Deep belief network, [23](#), [521](#), [622](#), [647](#), [650](#), [673](#), [681](#)
- Deep Blue, [2](#)
- Deep Boltzmann machine, [21](#), [23](#), [521](#), [622](#), [643](#), [647](#), [652](#), [661](#), [673](#)
- Deep feedforward network, [162](#), [416](#)
- Deep learning, [2](#), [5](#)
- Denoising autoencoder, [501](#), [678](#)
- Denoising score matching, [611](#)
- Density estimation, [100](#)
- Derivative, [xiii](#), [80](#)
- Design matrix, [103](#)
- Detector layer, [332](#)
- Determinant, [xii](#)
- Diagonal matrix, [38](#)
- Differential entropy, [71](#), [637](#)
- Dirac delta function, [62](#)
- Directed graphical model, [74](#), [498](#), [555](#), [681](#)
- Directional derivative, [82](#)
- Discriminative fine-tuning, *see* supervised fine-tuning
- Discriminative RBM, [676](#)
- Distributed representation, [16](#), [147](#), [538](#)
- Domain adaptation, [528](#)

- Dot product, 31, 137
- Double backprop, 267
- Doubly block circulant matrix, 326
- Dream sleep, 601, 643
- DropConnect, 261
- Dropout, 253, 416, 421, 422, 661, 678
- Dynamic structure, 439

- E-step, 625
- Early stopping, 239, 242–244, 416
- EBM, *see* energy-based model
- Echo state network, 21, 23, 395
- Effective capacity, 111
- Eigendecomposition, 39
- Eigenvalue, 39
- Eigenvector, 39
- ELBO, *see* evidence lower bound
- Element-wise product, *see* Hadamard product, *see* Hadamard product
- EM, *see* expectation maximization
- Embedding, 509
- Empirical distribution, 63
- Empirical risk, 270
- Empirical risk minimization, 270
- Encoder, 4
- Energy function, 562
- Energy-based model, 561, 587, 644, 653
- Ensemble methods, 250
- Epoch, 241
- Equality constraint, 91
- Equivariance, 331
- Error function, *see* objective function
- ESN, *see* echo state network
- Euclidean norm, 36
- Euler-Lagrange equation, 637
- Evidence lower bound, 624, 651
- Example, 96
- Expectation, 57
- Expectation maximization, 625
- Expected value, *see* expectation
- Explaining away, 566, 622, 635
- Exploitation, 471
- Exploration, 471
- Exponential distribution, 62

- F-score, 414
- Factor (graphical model), 559
- Factor analysis, 480
- Factor graph, 569
- Factors of variation, 4
- Feature, 96
- Feature selection, 230
- Feedforward neural network, 162
- Fine-tuning, 316
- Finite differences, 430
- Forget gate, 299
- Forward propagation, 198
- Fourier transform, 350, 353
- Fovea, 357
- FPCD, 606
- Free energy, 563, 669
- Freebase, 473
- Frequentist probability, 52
- Frequentist statistics, 132
- Frobenius norm, 43
- Fully-visible Bayes network, 694
- Functional derivatives, 636
- FVBN, *see* fully-visible Bayes network

- Gabor function, 359
- GANs, *see* generative adversarial networks
- Gated recurrent unit, 416
- Gaussian distribution, *see* normal distribution
- Gaussian kernel, 138
- Gaussian mixture, 64, 183
- GCN, *see* global contrast normalization
- GeneOntology, 473
- Generalization, 107
- Generalized Lagrange function, *see* generalized Lagrangian
- Generalized Lagrangian, 91
- Generative adversarial networks, 678, 688
- Generative moment matching networks, 692
- Generator network, 683
- Gibbs distribution, 560
- Gibbs sampling, 573, 591
- Global contrast normalization, 445
- GPU, *see* graphics processing unit
- Gradient, 81

- Gradient clipping, 282, 406
- Gradient descent, 80, 82
- Graph, xii
- Graphical model, *see* structured probabilistic model
- Graphics processing unit, 435
- Greedy algorithm, 316
- Greedy layer-wise unsupervised pretraining, 520
- Greedy supervised pretraining, 316
- Grid search, 423
- Hadamard product, xii, 31
- Hard tanh, 191
- Harmonium, *see* restricted Boltzmann machine
- Harmony theory, 563
- Helmholtz free energy, *see* evidence lower bound
- Hessian, 217
- Hessian matrix, xiii, 84
- Heteroscedastic, 182
- Hidden layer, 6, 162
- Hill climbing, 83
- Hyperparameter optimization, 423
- Hyperparameters, 117, 421
- Hypothesis space, 109, 115
- i.i.d. assumptions, 108, 119, 263
- Identity matrix, 33
- ILSVRC, *see* ImageNet Large Scale Visual Recognition Challenge
- ImageNet Large Scale Visual Recognition Challenge, 22
- Immortality, 569
- Importance sampling, 584, 615, 687
- Importance weighted autoencoder, 687
- Independence, xiii, 57
- Independent and identically distributed, *see* i.i.d. assumptions
- Independent component analysis, 481
- Independent subspace analysis, 483
- Inequality constraint, 91
- Inference, 554, 575, 622, 624, 626, 629, 639, 641
- Information retrieval, 516
- Initialization, 294
- Integral, xiii
- Invariance, 332
- Isotropic, 62
- Jacobian matrix, xiii, 69, 83
- Joint probability, 54
- k -means, 354, 540
- k -nearest neighbors, 139, 540
- Karush-Kuhn-Tucker conditions, 92, 231
- Karush-Kuhn-Tucker, 91
- Kernel (convolution), 324, 325
- Kernel machine, 540
- Kernel trick, 137
- KKT, *see* Karush-Kuhn-Tucker
- KKT conditions, *see* Karush-Kuhn-Tucker conditions
- KL divergence, *see* Kullback-Leibler divergence
- Knowledge base, 2, 473
- Krylov methods, 218
- Kullback-Leibler divergence, xiii, 71
- Label smoothing, 237
- Lagrange multipliers, 91, 637
- Lagrangian, *see* generalized Lagrangian
- LAPGAN, 691
- Laplace distribution, 62, 486
- Latent variable, 64
- Layer (neural network), 162
- LCN, *see* local contrast normalization
- Leaky ReLU, 187
- Leaky units, 398
- Learning rate, 82
- Line search, 82, 83, 90
- Linear combination, 34
- Linear dependence, 35
- Linear factor models, 479
- Linear regression, 104, 107, 136
- Link prediction, 474
- Lipschitz constant, 89
- Lipschitz continuous, 89
- Liquid state machine, 395

- Local conditional probability distribution, 556
- Local contrast normalization, 447
- Logistic regression, 3, 137, 137
- Logistic sigmoid, 7, 64
- Long short-term memory, 17, 24, 299, 400, 416
- Loop, 569
- Loopy belief propagation, 577
- Loss function, *see* objective function
- L^p norm, 36
- LSTM, *see* long short-term memory

- M-step, 625
- Machine learning, 2
- Machine translation, 98
- Main diagonal, 30
- Manifold, 156
- Manifold hypothesis, 157
- Manifold learning, 156
- Manifold tangent classifier, 267
- MAP approximation, 135, 496
- Marginal probability, 55
- Markov chain, 587
- Markov chain Monte Carlo, 587
- Markov network, *see* undirected model
- Markov random field, *see* undirected model
- Matrix, xi, xii, 29
- Matrix inverse, 33
- Matrix product, 31
- Max norm, 37
- Max pooling, 332
- Maximum likelihood, 128
- Maxout, 188, 416
- MCMC, *see* Markov chain Monte Carlo
- Mean field, 629, 630, 661
- Mean squared error, 105
- Measure theory, 68
- Measure zero, 68
- Memory network, 408
- Method of steepest descent, *see* gradient descent
- Minibatch, 273
- Missing inputs, 97
- Mixing (Markov chain), 593
- Mixture density networks, 183
- Mixture distribution, 63
- Mixture model, 183, 501
- Mixture of experts, 440, 540
- MLP, *see* multilayer perception
- MNIST, 19, 20, 661
- Model averaging, 250
- Model compression, 438
- Model identifiability, 278
- Model parallelism, 438
- Moment matching, 692
- Moore-Penrose pseudoinverse, 42, 234
- Moralized graph, 569
- MP-DBM, *see* multi-prediction DBM
- MRF (Markov Random Field), *see* undirected model
- MSE, *see* mean squared error
- Multi-modal learning, 532
- Multi-prediction DBM, 663
- Multi-task learning, 238, 530
- Multilayer perception, 5
- Multilayer perceptron, 23
- Multinomial distribution, 59
- Multinoulli distribution, 59

- n -gram, 452
- NADE, 697
- Naive Bayes, 3
- Nat, 70
- Natural image, 551
- Natural language processing, 451
- Nearest neighbor regression, 112
- Negative definite, 86
- Negative phase, 460, 598, 600
- Neocognitron, 15, 21, 23, 358
- Nesterov momentum, 293
- Netflix Grand Prize, 253, 470
- Neural language model, 454, 467
- Neural network, 13
- Neural Turing machine, 408
- Neuroscience, 14
- Newton's method, 86, 304
- NLM, *see* neural language model
- NLP, *see* natural language processing
- No free lunch theorem, 113

- Noise-contrastive estimation, 612
- Nonparametric model, 111
- Norm, xiv, 36
- Normal distribution, 60, 61, 122
- Normal equations, 106, 106, 109, 228
- Normalized initialization, 296
- Numerical differentiation, *see* finite differences

- Object detection, 443
- Object recognition, 443
- Objective function, 79
- OMP- k , *see* orthogonal matching pursuit
- One-shot learning, 530
- Operation, 199
- Optimization, 77, 79
- Orthodox statistics, *see* frequentist statistics
- Orthogonal matching pursuit, 23, 250
- Orthogonal matrix, 39
- Orthogonality, 38
- Output layer, 162

- Parallel distributed processing, 16
- Parameter initialization, 294, 397
- Parameter sharing, 247, 328, 364, 366, 379
- Parameter tying, *see* Parameter sharing
- Parametric model, 111
- Parametric ReLU, 187
- Partial derivative, 81
- Partition function, 560, 597, 659
- PCA, *see* principal components analysis
- PCD, *see* stochastic maximum likelihood
- Perceptron, 14, 23
- Persistent contrastive divergence, *see* stochastic maximum likelihood
- Perturbation analysis, *see* reparametrization trick
- Point estimator, 119
- Policy, 470
- Pooling, 323, 672
- Positive definite, 86
- Positive phase, 460, 598, 600, 646, 658
- Precision, 414
- Precision (of a normal distribution), 60, 62
- Predictive sparse decomposition, 515
- Preprocessing, 444
- Pretraining, 316, 520
- Primary visual cortex, 355
- Principal components analysis, 44, 143, 144, 480, 622
- Prior probability distribution, 132
- Probabilistic max pooling, 673
- Probabilistic PCA, 480, 481, 623
- Probability density function, 55
- Probability distribution, 53
- Probability mass function, 53
- Probability mass function estimation, 100
- Product of experts, 562
- Product rule of probability, *see* chain rule of probability
- PSD, *see* predictive sparse decomposition
- Pseudolikelihood, 607

- Quadrature pair, 360
- Quasi-Newton methods, 309

- Radial basis function, 191
- Random search, 425
- Random variable, 53
- Ratio matching, 610
- RBF, 191
- RBM, *see* restricted Boltzmann machine
- Recall, 414
- Receptive field, 329
- Recommender Systems, 468
- Rectified linear unit, 166, 187, 416, 498
- Recurrent network, 23
- Recurrent neural network, 369
- Regression, 97
- Regularization, 117, 117, 172, 222, 421
- Regularizer, 116
- REINFORCE, 679
- Reinforcement learning, 25, 103, 470, 678
- Relational database, 473
- Relations, 473
- Reparametrization trick, 678
- Representation learning, 3
- Representational capacity, 111
- Restricted Boltzmann machine, 347, 450, 470, 578, 622, 646, 647, 661, 666,

- 668, 670, 672
- Ridge regression, *see* weight decay
- Risk, 269
- RNN-RBM, 675
- Saddle points, 279
- Sample mean, 122
- Scalar, xi, xii, 28
- Score matching, 503, 609
- Second derivative, 83
- Second derivative test, 86
- Self-information, 70
- Semantic hashing, 516
- Semi-supervised learning, 238
- Separable convolution, 353
- Separation (probabilistic modeling), 564
- Set, xii
- SGD, *see* stochastic gradient descent
- Shannon entropy, xiii, 70
- Shortlist, 456
- Sigmoid, xiv, *see* logistic sigmoid
- Sigmoid belief network, 23
- Simple cell, 356
- Singular value, *see* singular value decomposition
- Singular value decomposition, 41, 144, 469
- Singular vector, *see* singular value decomposition
- Slow feature analysis, 483
- SML, *see* stochastic maximum likelihood
- Softmax, 178, 408, 440
- Softplus, xiv, 65, 191
- Spam detection, 3
- Sparse coding, 314, 347, 486, 622, 681
- Sparse initialization, 297, 397
- Sparse representation, 142, 220, 248, 496, 548
- Spearmint, 426
- Spectral radius, 395
- Speech recognition, *see* automatic speech recognition
- Sphering, *see* whitening
- Spike and slab restricted Boltzmann machine, 670
- SPN, *see* sum-product network
- Square matrix, 35
- ssRBM, *see* spike and slab restricted Boltzmann machine
- Standard deviation, 58
- Standard error, 124
- Standard error of the mean, 124, 272
- Statistic, 119
- Statistical learning theory, 107
- Steepest descent, *see* gradient descent
- Stochastic back-propagation, *see* reparametrization trick
- Stochastic gradient descent, 14, 147, 273, 287, 661
- Stochastic maximum likelihood, 604, 661
- Stochastic pooling, 261
- Structure learning, 574
- Structured output, 98, 674
- Structured probabilistic model, 74, 550
- Sum rule of probability, 55
- Sum-product network, 545
- Supervised fine-tuning, 521, 652
- Supervised learning, 102
- Support vector machine, 137
- Surrogate loss function, 270
- SVD, *see* singular value decomposition
- Symmetric matrix, 38, 40
- Tangent distance, 264
- Tangent plane, 507
- Tangent prop, 265
- TDNN, *see* time-delay neural network
- Teacher forcing, 374
- Tempering, 595
- Template matching, 138
- Tensor, xi, xii, 30
- Test set, 107
- Tikhonov regularization, *see* weight decay
- Tiled convolution, 343
- Time-delay neural network, 358, 365
- Toeplitz matrix, 326
- Topographic ICA, 483
- Trace operator, 43
- Training error, 107
- Transcription, 98
- Transfer learning, 528

- Transpose, [xii](#), [30](#)
- Triangle inequality, [36](#)
- Triangulated graph, *see* chordal graph
- Trigram, [452](#)

- Unbiased, [121](#)
- Undirected graphical model, [74](#), [498](#)
- Undirected model, [557](#)
- Uniform distribution, [54](#)
- Unigram, [452](#)
- Unit norm, [38](#)
- Unit vector, [38](#)
- Universal approximation theorem, [192](#)
- Universal approximator, [545](#)
- Unnormalized probability distribution, [559](#)
- Unsupervised learning, [102](#), [142](#)
- Unsupervised pretraining, [450](#), [520](#)

- V-structure, *see* explaining away
- V1, [355](#)
- VAE, *see* variational autoencoder
- Vapnik-Chervonenkis dimension, [111](#)
- Variance, [xiii](#), [58](#), [223](#)
- Variational autoencoder, [678](#), [685](#)
- Variational derivatives, *see* functional derivatives
- Variational free energy, *see* evidence lower bound
- VC dimension, *see* Vapnik-Chervonenkis dimension
- Vector, [xi](#), [xii](#), [29](#)
- Virtual adversarial examples, [264](#)
- Visible layer, [6](#)
- Volumetric data, [350](#)

- Wake-sleep, [642](#), [651](#)
- Weight decay, [115](#), [172](#), [225](#), [422](#)
- Weight space symmetry, [278](#)
- Weights, [14](#), [104](#)
- Whitening, [447](#)
- Wikibase, [473](#)
- Wikibase, [473](#)
- Word embedding, [454](#)
- Word-sense disambiguation, [475](#)
- WordNet, [473](#)

- Zero-data learning, *see* zero-shot learning
- Zero-shot learning, [530](#)