

Index

- 0-1 loss, [100](#), [271](#)
- Absolute value rectification, [187](#)
- Accuracy, [415](#)
- Activation function, [165](#)
- Active constraint, [92](#)
- AdaGrad, [301](#)
- ADALINE, *see* adaptive linear element
- Adam, [303](#), [417](#)
- Adaptive linear element, [14](#), [21](#), [23](#)
- Adversarial example, [263](#)
- Adversarial training, [264](#), [267](#), [523](#)
- Affine, [107](#)
- AIS, *see* annealed importance sampling
- Almost everywhere, [68](#)
- Almost sure convergence, [126](#)
- Ancestral sampling, [573](#), [588](#)
- ANN, *see* Artificial neural network
- Annealed importance sampling, [618](#), [659](#), [706](#)
- Approximate Bayesian computation, [706](#)
- Approximate inference, [576](#)
- Artificial intelligence, [1](#)
- Artificial neural network, *see* Neural network
- ASR, *see* automatic speech recognition
- Asymptotically unbiased, [121](#)
- Audio, [99](#), [351](#), [450](#)
- Autoencoder, [4](#), [348](#), [494](#)
- Automatic speech recognition, [450](#)
- Back-propagation, [198](#)
- Back-propagation through time, [376](#)
- Backprop, *see* back-propagation
- Bag of words, [462](#)
- Bagging, [250](#)
- Batch normalization, [262](#), [417](#)
- Bayes error, [114](#)
- Bayes' rule, [67](#)
- Bayesian hyperparameter optimization, [427](#)
- Bayesian network, *see* directed graphical model
- Bayesian probability, [52](#)
- Bayesian statistics, [132](#)
- Belief network, *see* directed graphical model
- Bernoulli distribution, [59](#)
- BFGS, [310](#)
- Bias, [121](#), [223](#)
- Bias parameter, [107](#)
- Biased importance sampling, [586](#)
- Bigram, [453](#)
- Binary relation, [474](#)
- Block Gibbs sampling, [592](#)
- Boltzmann distribution, [563](#)
- Boltzmann machine, [563](#), [645](#)
- BPTT, *see* back-propagation through time
- Broadcasting, [31](#)
- Burn-in, [590](#)
- CAE, *see* contractive autoencoder
- Calculus of variations, [173](#)
- Categorical distribution, *see* multinoulli distribution
- CD, *see* contrastive divergence
- Centering trick (DBM), [664](#)
- Central limit theorem, [61](#)
- Chain rule (calculus), [199](#)
- Chain rule of probability, [56](#)

- Chess, 2
- Chord, 570
- Chordal graph, 570
- Class-based language models, 455
- Classical dynamical system, 367
- Classification, 97
- Clique potential, *see* factor (graphical model)
- CNN, *see* convolutional neural network
- Collaborative Filtering, 470
- Collider, *see* explaining away
- Color images, 351
- Complex cell, 357
- Computational graph, 199
- Computer vision, 444
- Concept drift, 531
- Condition number, 274
- Conditional computation, *see* dynamic structure
- Conditional independence, xiv, 57
- Conditional probability, 56
- Conditional RBM, 676
- Connectionism, 16, 435
- Connectionist temporal classification, 452
- Consistency, 126, 504
- Constrained optimization, 90, 231
- Content-based addressing, 410
- Content-based recommender systems, 471
- Context-specific independence, 566
- Contextual bandits, 471
- Continuation methods, 321
- Contractive autoencoder, 513
- Contrast, 446
- Contrastive divergence, 285, 603, 662
- Convex optimization, 138
- Convolution, 324, 673
- Convolutional network, 15
- Convolutional neural network, 248, 324, 417, 451
- Coordinate descent, 315, 660
- Correlation, 58
- Cost function, *see* objective function
- Covariance, xiv, 58
- Covariance matrix, 59
- Coverage, 416
- Critical temperature, 596
- Cross-correlation, 326
- Cross-entropy, 72, 129
- Cross-validation, 119
- CTC, *see* connectionist temporal classification
- Curriculum learning, 322
- Curse of dimensionality, 151
- Cyc, 2
- D-separation, 565
- DAE, *see* denoising autoencoder
- Data generating distribution, 108, 128
- Data generating process, 108
- Data parallelism, 439
- Dataset, 101
- Dataset augmentation, 267, 449
- DBM, *see* deep Boltzmann machine
- DCGAN, 544, 545, 691
- Decision tree, 140, 541
- Decoder, 4
- Deep belief network, 23, 522, 623, 648, 651, 674, 682
- Deep Blue, 2
- Deep Boltzmann machine, 21, 23, 522, 623, 644, 648, 653, 662, 674
- Deep feedforward network, 162, 417
- Deep learning, 2, 5
- Denoising autoencoder, 502, 679
- Denoising score matching, 612
- Density estimation, 100
- Derivative, xiv, 80
- Design matrix, 103
- Detector layer, 333
- Determinant, xiii
- Diagonal matrix, 38
- Differential entropy, 71, 638
- Dirac delta function, 62
- Directed graphical model, 74, 499, 556, 682
- Directional derivative, 82
- Discriminative fine-tuning, *see* supervised fine-tuning
- Discriminative RBM, 677
- Distributed representation, 16, 147, 539
- Domain adaptation, 529

- Dot product, 31, 137
- Double backprop, 267
- Doubly block circulant matrix, 327
- Dream sleep, 602, 644
- DropConnect, 261
- Dropout, 253, 417, 422, 423, 662, 679
- Dynamic structure, 440

- E-step, 626
- Early stopping, 239, 242–244, 417
- EBM, *see* energy-based model
- Echo state network, 21, 23, 396
- 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, 510
- Empirical distribution, 63
- Empirical risk, 271
- Empirical risk minimization, 271
- Encoder, 4
- Energy function, 563
- Energy-based model, 562, 588, 645, 654
- Ensemble methods, 250
- Epoch, 241
- Equality constraint, 91
- Equivariance, 332
- Error function, *see* objective function
- ESN, *see* echo state network
- Euclidean norm, 36
- Euler-Lagrange equation, 638
- Evidence lower bound, 625, 652
- Example, 96
- Expectation, 57
- Expectation maximization, 626
- Expected value, *see* expectation
- Explaining away, 567, 623, 636
- Exploitation, 472
- Exploration, 472
- Exponential distribution, 62

- F-score, 415
- Factor (graphical model), 560
- Factor analysis, 481
- Factor graph, 570
- Factors of variation, 4
- Feature, 96
- Feature selection, 230
- Feedforward neural network, 162
- Fine-tuning, 317
- Finite differences, 431
- Forget gate, 300
- Forward propagation, 198
- Fourier transform, 351, 354
- Fovea, 358
- FPCD, 607
- Free energy, 564, 670
- Freebase, 474
- Frequentist probability, 52
- Frequentist statistics, 132
- Frobenius norm, 43
- Fully-visible Bayes network, 695
- Functional derivatives, 637
- FVBN, *see* fully-visible Bayes network

- Gabor function, 360
- GANs, *see* generative adversarial networks
- Gated recurrent unit, 417
- Gaussian distribution, *see* normal distribution
- Gaussian kernel, 138
- Gaussian mixture, 64, 183
- GCN, *see* global contrast normalization
- GeneOntology, 474
- Generalization, 107
- Generalized Lagrange function, *see* generalized Lagrangian
- Generalized Lagrangian, 91
- Generative adversarial networks, 679, 689
- Generative moment matching networks, 693
- Generator network, 684
- Gibbs distribution, 561
- Gibbs sampling, 574, 592
- Global contrast normalization, 446
- GPU, *see* graphics processing unit
- Gradient, 81

- Gradient clipping, 283, 407
- Gradient descent, 80, 82
- Graph, xiii
- Graphical model, *see* structured probabilistic model
- Graphics processing unit, 436
- Greedy algorithm, 317
- Greedy layer-wise unsupervised pretraining, 521
- Greedy supervised pretraining, 317
- Grid search, 424

- Hadamard product, xiii, 31
- Hard tanh, 191
- Harmonium, *see* restricted Boltzmann machine
- Harmony theory, 564
- Helmholtz free energy, *see* evidence lower bound
- Hessian, 217
- Hessian matrix, xiv, 84
- Heteroscedastic, 182
- Hidden layer, 6, 162
- Hill climbing, 83
- Hyperparameter optimization, 424
- Hyperparameters, 117, 422
- 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, 570
- Importance sampling, 585, 616, 688
- Importance weighted autoencoder, 688
- Independence, xiv, 57
- Independent and identically distributed, *see* i.i.d. assumptions
- Independent component analysis, 482
- Independent subspace analysis, 484
- Inequality constraint, 91
- Inference, 555, 576, 623, 625, 627, 630, 640, 642
- Information retrieval, 517
- Initialization, 295
- Integral, xiv
- Invariance, 333
- Isotropic, 62

- Jacobian matrix, xiv, 69, 83
- Joint probability, 54

- k -means, 355, 541
- k -nearest neighbors, 139, 541
- Karush-Kuhn-Tucker conditions, 92, 231
- Karush-Kuhn-Tucker, 91
- Kernel (convolution), 325, 326
- Kernel machine, 541
- Kernel trick, 137
- KKT, *see* Karush-Kuhn-Tucker
- KKT conditions, *see* Karush-Kuhn-Tucker conditions
- KL divergence, *see* Kullback-Leibler divergence
- Knowledge base, 2, 474
- Krylov methods, 218
- Kullback-Leibler divergence, xiv, 71

- Label smoothing, 237
- Lagrange multipliers, 91, 638
- Lagrangian, *see* generalized Lagrangian
- LAPGAN, 692
- Laplace distribution, 62, 487
- Latent variable, 64
- Layer (neural network), 162
- LCN, *see* local contrast normalization
- Leaky ReLU, 187
- Leaky units, 399
- Learning rate, 82
- Line search, 82, 83, 90
- Linear combination, 34
- Linear dependence, 35
- Linear factor models, 480
- Linear regression, 104, 107, 136
- Link prediction, 475
- Lipschitz constant, 89
- Lipschitz continuous, 89
- Liquid state machine, 396

- Local conditional probability distribution, 557
- Local contrast normalization, 448
- Logistic regression, 3, 137, 137
- Logistic sigmoid, 7, 64
- Long short-term memory, 17, 24, 300, 401, 417
- Loop, 570
- Loopy belief propagation, 578
- Loss function, *see* objective function
- L^p norm, 36
- LSTM, *see* long short-term memory
- M-step, 626
- Machine learning, 2
- Machine translation, 98
- Main diagonal, 30
- Manifold, 156
- Manifold hypothesis, 157
- Manifold learning, 156
- Manifold tangent classifier, 267
- MAP approximation, 135, 497
- Marginal probability, 55
- Markov chain, 588
- Markov chain Monte Carlo, 588
- Markov network, *see* undirected model
- Markov random field, *see* undirected model
- Matrix, xii, xiii, 29
- Matrix inverse, 33
- Matrix product, 31
- Max norm, 37
- Max pooling, 333
- Maximum likelihood, 128
- Maxout, 188, 417
- MCMC, *see* Markov chain Monte Carlo
- Mean field, 630, 631, 662
- Mean squared error, 105
- Measure theory, 68
- Measure zero, 68
- Memory network, 409
- Method of steepest descent, *see* gradient descent
- Minibatch, 274
- Missing inputs, 97
- Mixing (Markov chain), 594
- Mixture density networks, 183
- Mixture distribution, 63
- Mixture model, 183, 502
- Mixture of experts, 441, 541
- MLP, *see* multilayer perception
- MNIST, 19, 20, 662
- Model averaging, 250
- Model compression, 439
- Model identifiability, 279
- Model parallelism, 439
- Moment matching, 693
- Moore-Penrose pseudoinverse, 42, 234
- Moralized graph, 570
- MP-DBM, *see* multi-prediction DBM
- MRF (Markov Random Field), *see* undirected model
- MSE, *see* mean squared error
- Multi-modal learning, 533
- Multi-prediction DBM, 664
- Multi-task learning, 238, 531
- Multilayer perception, 5
- Multilayer perceptron, 23
- Multinomial distribution, 59
- Multinoulli distribution, 59
- n -gram, 453
- NADE, 698
- Naive Bayes, 3
- Nat, 70
- Natural image, 552
- Natural language processing, 452
- Nearest neighbor regression, 112
- Negative definite, 86
- Negative phase, 461, 599, 601
- Neocognitron, 15, 21, 23, 359
- Nesterov momentum, 294
- Netflix Grand Prize, 253, 471
- Neural language model, 455, 468
- Neural network, 13
- Neural Turing machine, 409
- Neuroscience, 14
- Newton's method, 86, 305
- NLM, *see* neural language model
- NLP, *see* natural language processing
- No free lunch theorem, 113

- Noise-contrastive estimation, 613
- Nonparametric model, 111
- Norm, xv, 36
- Normal distribution, 60, 61, 122
- Normal equations, 106, 106, 109, 228
- Normalized initialization, 297
- Numerical differentiation, *see* finite differences

- Object detection, 444
- Object recognition, 444
- Objective function, 79
- OMP- k , *see* orthogonal matching pursuit
- One-shot learning, 531
- 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, 295, 398
- Parameter sharing, 247, 329, 365, 367, 380
- Parameter tying, *see* Parameter sharing
- Parametric model, 111
- Parametric ReLU, 187
- Partial derivative, 81
- Partition function, 561, 598, 660
- 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, 471
- Pooling, 324, 673
- Positive definite, 86
- Positive phase, 461, 599, 601, 647, 659
- Precision, 415
- Precision (of a normal distribution), 60, 62
- Predictive sparse decomposition, 516

- Preprocessing, 445
- Pretraining, 317, 521
- Primary visual cortex, 356
- Principal components analysis, 44, 143, 144, 481, 623
- Prior probability distribution, 132
- Probabilistic max pooling, 674
- Probabilistic PCA, 481, 482, 624
- Probability density function, 55
- Probability distribution, 53
- Probability mass function, 53
- Probability mass function estimation, 100
- Product of experts, 563
- Product rule of probability, *see* chain rule of probability
- PSD, *see* predictive sparse decomposition
- Pseudolikelihood, 608

- Quadrature pair, 361
- Quasi-Newton methods, 310

- Radial basis function, 191
- Random search, 426
- Random variable, 53
- Ratio matching, 611
- RBF, 191
- RBM, *see* restricted Boltzmann machine
- Recall, 415
- Receptive field, 330
- Recommender Systems, 469
- Rectified linear unit, 166, 187, 417, 499
- Recurrent network, 23
- Recurrent neural network, 370
- Regression, 97
- Regularization, 117, 117, 172, 222, 422
- Regularizer, 116
- REINFORCE, 680
- Reinforcement learning, 25, 103, 471, 679
- Relational database, 474
- Relations, 474
- Reparametrization trick, 679
- Representation learning, 3
- Representational capacity, 111
- Restricted Boltzmann machine, 348, 451, 471, 579, 623, 647, 648, 662, 667,

- 669, 671, 673
- Ridge regression, *see* weight decay
- Risk, 270
- RNN-RBM, 676
- Saddle points, 280
- Sample mean, 122
- Scalar, xii, xiii, 28
- Score matching, 504, 610
- Second derivative, 83
- Second derivative test, 86
- Self-information, 70
- Semantic hashing, 517
- Semi-supervised learning, 238
- Separable convolution, 354
- Separation (probabilistic modeling), 565
- Set, xiii
- SGD, *see* stochastic gradient descent
- Shannon entropy, xiv, 70
- Shortlist, 457
- Sigmoid, xv, *see* logistic sigmoid
- Sigmoid belief network, 23
- Simple cell, 357
- Singular value, *see* singular value decomposition
- Singular value decomposition, 41, 144, 470
- Singular vector, *see* singular value decomposition
- Slow feature analysis, 484
- SML, *see* stochastic maximum likelihood
- Softmax, 178, 409, 441
- Softplus, xv, 65, 191
- Spam detection, 3
- Sparse coding, 315, 348, 487, 623, 682
- Sparse initialization, 298, 398
- Sparse representation, 142, 220, 248, 497, 549
- Spearmint, 427
- Spectral radius, 396
- Speech recognition, *see* automatic speech recognition
- Sphering, *see* whitening
- Spike and slab restricted Boltzmann machine, 671
- 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, 273
- Statistic, 119
- Statistical learning theory, 107
- Steepest descent, *see* gradient descent
- Stochastic back-propagation, *see* reparametrization trick
- Stochastic gradient descent, 14, 147, 274, 288, 662
- Stochastic maximum likelihood, 605, 662
- Stochastic pooling, 261
- Structure learning, 575
- Structured output, 98, 675
- Structured probabilistic model, 74, 551
- Sum rule of probability, 55
- Sum-product network, 546
- Supervised fine-tuning, 522, 653
- Supervised learning, 102
- Support vector machine, 137
- Surrogate loss function, 271
- SVD, *see* singular value decomposition
- Symmetric matrix, 38, 40
- Tangent distance, 265
- Tangent plane, 508
- Tangent prop, 265
- TDNN, *see* time-delay neural network
- Teacher forcing, 375
- Tempering, 596
- Template matching, 138
- Tensor, xii, xiii, 30
- Test set, 107
- Tikhonov regularization, *see* weight decay
- Tiled convolution, 344
- Time-delay neural network, 359, 366
- Toeplitz matrix, 327
- Topographic ICA, 484
- Trace operator, 43
- Training error, 107
- Transcription, 98
- Transfer learning, 529

- Transpose, [xiii](#), [30](#)
- Triangle inequality, [36](#)
- Triangulated graph, *see* chordal graph
- Trigram, [453](#)

- Unbiased, [121](#)
- Undirected graphical model, [74](#), [499](#)
- Undirected model, [558](#)
- Uniform distribution, [54](#)
- Unigram, [453](#)
- Unit norm, [38](#)
- Unit vector, [38](#)
- Universal approximation theorem, [192](#)
- Universal approximator, [546](#)
- Unnormalized probability distribution, [560](#)
- Unsupervised learning, [102](#), [142](#)
- Unsupervised pretraining, [451](#), [521](#)

- V-structure, *see* explaining away
- V1, [356](#)
- VAE, *see* variational autoencoder
- Vapnik-Chervonenkis dimension, [111](#)
- Variance, [xiv](#), [58](#), [223](#)
- Variational autoencoder, [679](#), [686](#)
- Variational derivatives, *see* functional derivatives
- Variational free energy, *see* evidence lower bound
- VC dimension, *see* Vapnik-Chervonenkis dimension
- Vector, [xii](#), [xiii](#), [29](#)
- Virtual adversarial examples, [264](#)
- Visible layer, [6](#)
- Volumetric data, [351](#)

- Wake-sleep, [643](#), [652](#)
- Weight decay, [115](#), [172](#), [225](#), [423](#)
- Weight space symmetry, [279](#)
- Weights, [14](#), [104](#)
- Whitening, [448](#)
- Wikibase, [474](#)
- Wikibase, [474](#)
- Word embedding, [455](#)
- Word-sense disambiguation, [476](#)
- WordNet, [474](#)

- Zero-data learning, *see* zero-shot learning
- Zero-shot learning, [531](#)