

# Deep Learning

Yoshua Bengio  
Ian J. Goodfellow  
Aaron Courville

March 30, 2015

# Table of Contents

<b>Acknowledgments</b>	<b>1</b>
<b>Notation</b>	<b>2</b>
<b>1 Introduction</b>	<b>4</b>
1.1 Who Should Read This Book? . . . . .	10
1.2 Historical Trends in Deep Learning . . . . .	12
<b>I Applied math and machine learning basics</b>	<b>18</b>
<b>2 Linear Algebra</b>	<b>20</b>
2.1 Scalars, Vectors, Matrices and Tensors . . . . .	20
2.2 Multiplying Matrices and Vectors . . . . .	22
2.3 Identity and Inverse Matrices . . . . .	24
2.4 Linear Dependence, Span, and Rank . . . . .	25
2.5 Norms . . . . .	26
2.6 Special Kinds of Matrices and Vectors . . . . .	28
2.7 Eigendecomposition . . . . .	29
2.8 Singular Value Decomposition . . . . .	30
2.9 The Trace Operator . . . . .	31
2.10 Determinant . . . . .	31
2.11 Example: Principal Components Analysis . . . . .	32
<b>3 Probability and Information Theory</b>	<b>35</b>
3.1 Why Probability? . . . . .	35
3.2 Random Variables . . . . .	37
3.3 Probability Distributions . . . . .	37
3.3.1 Discrete Variables and Probability Mass Functions . . . . .	38
3.3.2 Continuous Variables and Probability Density Functions . . . . .	38
3.4 Marginal Probability . . . . .	39
3.5 Conditional Probability . . . . .	39
3.6 The Chain Rule of Conditional Probabilities . . . . .	40
3.7 Independence and Conditional Independence . . . . .	40

3.8	Expectation, Variance, and Covariance	41
3.9	Information Theory	42
3.10	Common Probability Distributions	44
3.10.1	Bernoulli Distribution	44
3.10.2	Multinoulli Distribution	44
3.10.3	Gaussian Distribution	45
3.10.4	Dirac Distribution	47
3.10.5	Mixtures of Distributions and Gaussian Mixture	48
3.11	Useful Properties of Common Functions	48
3.12	Bayes' Rule	51
3.13	Technical Details of Continuous Variables	51
3.14	Example: Naive Bayes	52
<b>4</b>	<b>Numerical Computation</b>	<b>56</b>
4.1	Overflow and Underflow	56
4.2	Poor Conditioning	57
4.3	Gradient-Based Optimization	58
4.4	Constrained Optimization	65
4.5	Example: Linear Least Squares	68
<b>5</b>	<b>Machine Learning Basics</b>	<b>70</b>
5.1	Learning Algorithms	70
5.1.1	The Task, $T$	70
5.1.2	The Performance Measure, $P$	72
5.1.3	The Experience, $E$	73
5.2	Example: Linear Regression	74
5.3	Generalization, Capacity, Overfitting and Underfitting	76
5.3.1	Generalization	76
5.3.2	Capacity	77
5.3.3	Occam's Razor, Underfitting and Overfitting	78
5.4	Estimating and Monitoring Generalization Error	81
5.5	Estimators, Bias, and Variance	83
5.5.1	Point Estimation	83
5.5.2	Bias	84
5.5.3	Variance	85
5.5.4	Trading off Bias and Variance and the Mean Squared Error	85
5.5.5	Consistency	86
5.6	Maximum Likelihood Estimation	87
5.6.1	Properties of Maximum Likelihood	87
5.6.2	Regularized Likelihood	87
5.7	Bayesian Statistics	87
5.8	Supervised Learning	88
5.8.1	Estimating Conditional Expectation by Minimizing Squared Error	88

5.8.2	Estimating Probabilities or Conditional Probabilities by Maximum Likelihood . . . . .	89
5.9	Unsupervised Learning . . . . .	90
5.9.1	Principal Components Analysis . . . . .	91
5.10	Weakly Supervised Learning . . . . .	93
5.11	The Smoothness Prior, Local Generalization and Non-Parametric Models . . . . .	93
5.12	Manifold Learning and the Curse of Dimensionality . . . . .	97
5.13	Challenges of High-Dimensional Distributions . . . . .	100
<b>II</b>	<b>Modern practical deep networks</b>	<b>102</b>
<b>6</b>	<b>Feedforward Deep Networks</b>	<b>104</b>
6.1	Formalizing and Generalizing Neural Networks . . . . .	104
6.2	Parametrizing a Learned Predictor . . . . .	108
6.2.1	Family of Functions . . . . .	108
6.2.2	Loss Function and Conditional Log-Likelihood . . . . .	109
6.2.3	Training Criterion and Regularizer . . . . .	115
6.2.4	Optimization Procedure . . . . .	116
6.3	Flow Graphs and Back-Propagation . . . . .	117
6.3.1	Chain Rule . . . . .	118
6.3.2	Back-Propagation in an MLP . . . . .	119
6.3.3	Back-Propagation in a General Flow Graph . . . . .	120
6.4	Universal Approximation Properties and Depth . . . . .	126
6.5	Feature / Representation Learning . . . . .	128
6.6	Piecewise Linear Hidden Units . . . . .	129
6.7	Historical Notes . . . . .	130
<b>7</b>	<b>Regularization</b>	<b>131</b>
7.1	Classical Regularization: Parameter Norm Penalty . . . . .	132
7.1.1	$L^2$ Parameter Regularization . . . . .	133
7.1.2	$L^1$ Regularization . . . . .	135
7.1.3	$L^\infty$ Regularization . . . . .	138
7.2	Classical Regularization as Constrained Optimization . . . . .	138
7.3	Regularization from a Bayesian Perspective . . . . .	139
7.4	Regularization and Under-Constrained Problems . . . . .	139
7.5	Dataset Augmentation . . . . .	141
7.6	Classical Regularization as Noise Robustness . . . . .	142
7.7	Bagging and Other Ensemble Methods . . . . .	142
7.8	Early Stopping as a Form of Regularization . . . . .	143
7.9	Parameter Sharing . . . . .	150
7.10	Sparse Representations . . . . .	151
7.11	Dropout . . . . .	151
7.12	Multi-Task Learning . . . . .	154

<b>8</b>	<b>Optimization for Training Deep Models</b>	<b>156</b>
8.1	Optimization for Model Training	156
8.1.1	Empirical Risk Minimization	156
8.1.2	Surrogate Loss Functions	157
8.1.3	Generalization	157
8.1.4	Batches and Minibatches	158
8.1.5	Data Parallelism	158
8.2	Challenges in Optimization	158
8.2.1	Local Minima	158
8.2.2	Ill-Conditioning	158
8.2.3	Plateaus, Saddle Points, and Other Flat Regions	158
8.2.4	Cliffs and Exploding Gradients	158
8.2.5	Vanishing and Exploding Gradients - An Introduction to the Issue of Learning Long-Term Dependencies	161
8.3	Optimization Algorithms	164
8.3.1	Gradient Descent	164
8.3.2	Stochastic Gradient Descent	165
8.3.3	Momentum	166
8.3.4	Adagrad	167
8.3.5	RMSprop	167
8.3.6	Adadelat	168
8.3.7	No Pesky Learning Rates	168
8.4	Approximate Natural Gradient and Second-Order Methods	168
8.5	Conjugate Gradients	168
8.6	BFGS	168
8.6.1	New	168
8.6.2	Optimization Strategies and Meta-Algorithms	168
8.6.3	Coordinate Descent	168
8.6.4	Greedy Supervised Pre-training	169
8.7	Hints and Curriculum Learning	169
<b>9</b>	<b>Convolutional Networks</b>	<b>173</b>
9.1	The Convolution Operation	173
9.2	Motivation	175
9.3	Pooling	178
9.4	Variants of the Basic Convolution Function	183
9.5	Data Types	188
9.6	Efficient Convolution Algorithms	190
9.7	Deep Learning History	190
<b>10</b>	<b>Sequence Modeling: Recurrent and Recursive Nets</b>	<b>191</b>
10.1	Unfolding Flow Graphs and Sharing Parameters	191
10.2	Recurrent Neural Networks	193
10.2.1	Computing the Gradient in a Recurrent Neural Network	195

10.2.2	Recurrent Networks as Generative Directed Acyclic Models . . .	197
10.2.3	RNNs to Represent Conditional Probability Distributions . . . .	199
10.3	Bidirectional RNNs . . . . .	201
10.4	Recursive Neural Networks . . . . .	204
10.5	Auto-Regressive Networks . . . . .	205
10.5.1	Logistic Auto-Regressive Networks . . . . .	206
10.5.2	Neural Auto-Regressive Networks . . . . .	207
10.5.3	NADE . . . . .	209
10.6	Facing the Challenge of Long-Term Dependencies . . . . .	210
10.6.1	Echo State Networks: Choosing Weights to Make Dynamics Barely Contractive . . . . .	210
10.6.2	Combining Short and Long Paths in the Unfolded Flow Graph .	212
10.6.3	Leaky Units and a Hierarchy of Different Time Scales . . . . .	213
10.6.4	The Long-Short-Term-Memory Architecture and Other Gated RNNs	214
10.6.5	Deep RNNs . . . . .	217
10.6.6	Better Optimization . . . . .	218
10.6.7	Clipping Gradients . . . . .	219
10.6.8	Regularizing to Encourage Information Flow . . . . .	220
10.6.9	Organizing the State at Multiple Time Scales . . . . .	221
10.7	Handling Temporal Dependencies with N-Grams, HMMs, CRFs and Other Graphical Models . . . . .	222
10.7.1	N-grams . . . . .	222
10.7.2	Efficient Marginalization and Inference for Temporally Structured Outputs by Dynamic Programming . . . . .	223
10.7.3	HMMs . . . . .	227
10.7.4	CRFs . . . . .	229
10.8	Combining Neural Networks and Search . . . . .	231
10.8.1	Approximate Search . . . . .	233
<b>11</b>	<b>Large scale deep learning</b>	<b>236</b>
11.1	Fast CPU Implementations . . . . .	236
11.2	GPU Implementations . . . . .	236
11.3	Asynchronous Parallel Implementations . . . . .	236
11.4	Model Compression . . . . .	236
11.5	Dynamically Structured Nets . . . . .	237
<b>12</b>	<b>Practical methodology</b>	<b>238</b>
12.1	When to Gather More Data, Control Capacity, or Change Algorithms .	238
12.2	Machine Learning Methodology 101 . . . . .	238
12.3	Manual Hyperparameter Tuning . . . . .	238
12.4	Hyper-parameter Optimization Algorithms . . . . .	238
12.5	Tricks of the Trade for Deep Learning . . . . .	240
12.5.1	Debugging Back-Prop . . . . .	240

12.5.2	Automatic Differentiation and Symbolic Manipulations of Flow Graphs . . . . .	240
12.5.3	Momentum and Other Averaging Techniques as Cheap Second Order Methods . . . . .	240
<b>13</b>	<b>Applications</b>	<b>241</b>
13.1	Computer Vision . . . . .	241
13.1.1	Preprocessing . . . . .	242
13.1.2	Convolutional Nets . . . . .	247
13.2	Speech Recognition . . . . .	247
13.3	Natural Language Processing and Neural Language Models . . . . .	248
13.3.1	The Basics of Neural Language Models . . . . .	248
13.3.2	The Problem With N-Grams . . . . .	248
13.3.3	How Neural Language Models can Generalize Better . . . . .	250
13.3.4	Neural Machine Translation . . . . .	252
13.3.5	High-Dimensional Outputs . . . . .	252
13.3.6	Combining Neural Language Models with N-Grams . . . . .	252
13.4	Structured Outputs . . . . .	252
13.5	Other Applications . . . . .	252
<b>III</b>	<b>Deep learning research</b>	<b>253</b>
<b>14</b>	<b>Structured Probabilistic Models: A Deep Learning Perspective</b>	<b>255</b>
14.1	The Challenge of Unstructured Modeling . . . . .	256
14.2	Using Graphs to Describe Model Structure . . . . .	259
14.2.1	Directed Models . . . . .	259
14.2.2	Undirected Models . . . . .	261
14.2.3	The Partition Function . . . . .	263
14.2.4	Energy-Based Models . . . . .	265
14.2.5	Separation and D-Separation . . . . .	266
14.2.6	Converting Between Undirected and Directed Graphs . . . . .	267
14.2.7	Marginalizing Variables out of a Graph . . . . .	272
14.2.8	Factor Graphs . . . . .	272
14.3	Advantages of Structured Modeling . . . . .	272
14.4	Learning About Dependencies . . . . .	273
14.4.1	Latent Variables Versus Structure Learning . . . . .	273
14.4.2	Latent Variables for Feature Learning . . . . .	274
14.5	Inference and Approximate Inference Over Latent Variables . . . . .	274
14.5.1	Reparametrization Trick . . . . .	274
14.6	The Deep Learning Approach to Structured Probabilistic Modeling . . . . .	276
14.6.1	Example: The Restricted Boltzmann Machine . . . . .	277

<b>15 Monte Carlo Methods</b>	<b>279</b>
15.1 Markov Chain Monte Carlo Methods	279
15.1.1 Markov Chain Theory	280
<b>16 Linear Factor Models and Auto-Encoders</b>	<b>282</b>
16.1 Regularized Auto-Encoders	283
16.2 Representational Power, Layer Size and Depth	287
16.3 Reconstruction Distribution	288
16.4 Linear Factor Models	289
16.5 Probabilistic PCA and Factor Analysis	289
16.5.1 ICA	291
16.5.2 Sparse Coding as a Generative Model	292
16.6 Probabilistic Interpretation of Reconstruction Error as Log-Likelihood	293
16.7 Sparse Representations	295
16.7.1 Sparse Auto-Encoders	296
16.7.2 Predictive Sparse Decomposition	298
16.8 Denoising Auto-Encoders	298
16.8.1 Learning a Vector Field that Estimates a Gradient Field	301
16.9 Contractive Auto-Encoders	304
<b>17 Representation Learning</b>	<b>307</b>
17.1 Greedy Layerwise Unsupervised Pre-Training	308
17.1.1 Why Does Unsupervised Pre-Training Work?	311
17.2 Transfer Learning and Domain Adaptation	315
17.3 Semi-Supervised Learning	322
17.4 Causality, Semi-Supervised Learning and Disentangling the Underlying Factors	323
17.5 Assumption of Underlying Factors and Distributed Representation	325
17.6 Exponential Gain in Representational Efficiency from Distributed Representations	329
17.7 Exponential Gain in Representational Efficiency from Depth	330
17.8 Priors Regarding The Underlying Factors	333
<b>18 The Manifold Perspective on Representation Learning</b>	<b>336</b>
18.1 Manifold Interpretation of PCA and Linear Auto-Encoders	344
18.2 Manifold Interpretation of Sparse Coding	347
18.3 Manifold Learning via Regularized Auto-Encoders	347
18.4 Tangent Distance, Tangent-Prop, and Manifold Tangent Classifier	348
<b>19 Confronting the Partition Function</b>	<b>352</b>
19.1 Estimating the Partition Function	352
19.1.1 Annealed Importance Sampling	354
19.1.2 Bridge Sampling	357
19.1.3 Extensions	357



19.2	Stochastic Maximum Likelihood and Contrastive Divergence . . . . .	358
19.3	Pseudolikelihood . . . . .	365
19.4	Score Matching and Ratio Matching . . . . .	367
19.5	Denoising Score Matching . . . . .	369
19.6	Noise-Contrastive Estimation . . . . .	370
<b>20</b>	<b>Approximate inference</b>	<b>372</b>
20.1	Inference as Optimization . . . . .	372
20.2	Expectation Maximization . . . . .	375
20.3	MAP Inference: Sparse Coding as a Probabilistic Model . . . . .	375
20.4	Variational Inference and Learning . . . . .	376
20.4.1	Discrete Latent Variables . . . . .	378
20.4.2	Calculus of Variations . . . . .	378
20.4.3	Continuous Latent Variables . . . . .	380
20.5	Stochastic Inference . . . . .	380
20.6	Learned Approximate Inference . . . . .	380
<b>21</b>	<b>Deep Generative Models</b>	<b>382</b>
21.1	Restricted Boltzmann Machines . . . . .	382
21.1.1	Conditional Distributions . . . . .	384
21.1.2	RBM Gibbs Sampling . . . . .	385
21.2	Training Restricted Boltzmann Machines . . . . .	385
21.2.1	Contrastive Divergence Training of the RBM . . . . .	388
21.2.2	Stochastic Maximum Likelihood (Persistent Contrastive Divergence) for the RBM . . . . .	388
21.2.3	Other Inductive Principles . . . . .	388
21.3	Deep Belief Networks . . . . .	389
21.4	Deep Boltzmann Machines . . . . .	391
21.4.1	Interesting Properties . . . . .	395
21.4.2	Mean Field Inference in the DBM . . . . .	395
21.4.3	Variational Expectation Maximization . . . . .	396
21.4.4	Variational Learning With SML . . . . .	396
21.4.5	Layerwise Pretraining . . . . .	397
21.4.6	Multi-Prediction Deep Boltzmann Machines . . . . .	398
21.4.7	Centered Deep Boltzmann Machines . . . . .	398
21.5	Boltzmann Machines for Real-Valued Data . . . . .	400
21.5.1	Gaussian-Bernoulli RBMs . . . . .	400
21.5.2	mcRBMs . . . . .	400
21.5.3	mPoT Model . . . . .	400
21.5.4	Spike and Slab Restricted Boltzmann Machines . . . . .	401
21.6	Convolutional Boltzmann Machines . . . . .	401
21.7	Other Boltzmann Machines . . . . .	402
21.8	Directed Generative Nets . . . . .	403
21.8.1	Sigmoid Belief Nets . . . . .	403

21.8.2 Variational Autoencoders . . . . .	403
21.8.3 Variational Interpretation of PSD . . . . .	403
21.8.4 Generative Adversarial Networks . . . . .	403
21.9 A Generative View of Autoencoders . . . . .	404
21.9.1 Markov Chain Associated with any Denoising Auto-Encoder . . . . .	405
21.9.2 Clamping and Conditional Sampling . . . . .	407
21.9.3 Walk-Back Training Procedure . . . . .	408
21.10 Generative Stochastic Networks . . . . .	409
21.10.1 Discriminant GSNs . . . . .	410
21.11 Methodological Notes . . . . .	411
<b>Bibliography</b>	<b>413</b>
<b>Index</b>	<b>437</b>

# Acknowledgments

We would like to thank the following people who commented our proposal for the book and helped plan its contents and organization: Hugo Larochelle, Guillaume Alain, Kyunghyun Cho, Caglar Gulcehre (TODO diacritics), Razvan Pascanu, David Krueger and Thomas Rohée.

We would like to thank the following people who offered feedback on the content of the book itself:

In many chapters: Julian Serban, Laurent Dinh, Guillaume Alain, Ilya Sutskever, Vincent Vanhoucke, David Warde-Farley, Jurgen Van Gael, Dustin Webb, Johannes Roith, Ion Androutsopoulos, Pawel Chilinski, Halis Sak, Grigory Sapunov, Ion Androutsopoulos.

Introduction: Johannes Roith, Eric Morris, Samira Ebrahimi, Ozan Çaglayan, Martín Abadi.

Math background chapters:

Linear algebra: Pierre Luc Carrier, Li Yao, Thomas Rohée, Colby Toland, Amjad Almahairi, Sergey Oreshkov,

Probability: Rasmus Antti, Stephan Gouws, Vincent Dumoulin, Artem Oboturov, Li Yao, John Philip Anderson.

Numerical: Meire Fortunato,

Optimization: Marcel Ackermann

ML: Dzmitry Bahdanau Kelvin Xu

MLPs:

Convolutional nets: Mehdi Mirza, Caglar Gulcehre.

Unsupervised: Kelvin Xu

Partition function: Sam Bowman.

Graphical models: Kelvin Xu

RNNs: Kelvin Xu Dmitriy Serdyuk Dongyu Shi

We also want to thank David Warde-Farley, Matthew D. Zeiler, Rob Fergus, Chris Olah, Jason Yosinski, Nicolas Chapados and James Bergstra for contributing images or figures (as noted in the captions).

TODO– this section is just notes, write it up in nice presentation form.

# Notation

## Mathematical Objects

$a$	A scalar (integer or real) value with the name “a”
$\mathbf{a}$	A vector with the name “a”
$\mathbf{A}$	A matrix with the name “A”
TODO	TODO– higher order tensors
$\mathbb{A}$	A set with the name “A”
$\mathbb{R}$	The set of real numbers
$\{0, 1\}$	The set containing 0 and 1
$\mathbf{a}$	A scalar random variable with the name “a”
$\mathbf{a}$	A vector-valued random variable with the name “a”
$\mathbf{A}$	A matrix-valued random variable with the name “A”
$\mathcal{G}$	A graph with the name “G”

## Indexing

$a_i$	Element $i$ of vector $\mathbf{a}$ , with indexing starting at 1
$A_{i,j}$	Element $i, j$ of matrix $\mathbf{A}$
$\mathbf{A}_{i,:}$	Row $i$ of matrix $\mathbf{A}$
$\mathbf{A}_{:,i}$	Column $i$ of matrix $\mathbf{A}$
TODO	TODO– higher order tensors
$\mathbf{a}_i$	Element $i$ of the random vector $\mathbf{a}$
$\mathbf{x}^{(t)}$	usually the $t$ -th example (input) from a dataset, with $y^{(t)}$ the associated target, for supervised learning
$\mathbf{X}$	The matrix of input examples, with one row per example $\mathbf{x}^{(t)}$ .

## Linear Algebra Operations

$\mathbf{A}^\top$  Transpose of matrix  $\mathbf{A}$

$\mathbf{A} \odot \mathbf{B}$  Element-wise (Hadamard) product of  $\mathbf{A}$  and  $\mathbf{B}$

## Calculus

$\frac{dy}{dx}$  Derivative of  $y$  with respect to  $x$

$\frac{\partial y}{\partial x}$  Partial derivative of  $y$  with respect to  $x$

$\nabla_{\mathbf{x}} y$  Gradient of  $y$  with respect to  $\mathbf{x}$

$\nabla_{\mathbf{X}} y$  Matrix derivatives of  $y$  with respect to  $\mathbf{x}$

$\int f(\mathbf{x}) d\mathbf{x}$  Definite integral over the entire domain of  $\mathbf{x}$

$\int_{\mathbb{S}} f(\mathbf{x}) d\mathbf{x}$  Definite integral with respect to  $\mathbf{x}$  over the set  $\mathbb{S}$

## Miscellaneous

$f \circ g$  Composition of the functions  $f$  and  $g$

$\log x$  Natural logarithm of  $x$

## Probability and Information Theory

$a \perp b$  The random variables  $a$  and  $b$  are independent.

$a \perp b \mid c$  The random variables  $a$  and  $b$  are conditionally independent given  $c$ .

$\mathbb{E}_{x \sim P}[f(x)]$  or  $\mathbb{E}f(x)$  Expectation of  $f(x)$  with respect to  $P(\mathbf{x})$

$Var(f(x))$  Variance of  $f(x)$  under  $P(\mathbf{x})$

$Cov(f(x), g(x))$  Covariance of  $f(x)$  and  $g(x)$  under  $P(\mathbf{x}, \mathbf{y})$

$D_{\text{KL}}(P \parallel Q)$  Kullback-Leibler divergence of  $P$  and  $Q$

TODO– norms TODO– entropy TODO– Jacobian and Hessian TODO– Specify that unless otherwise clear from context, functions applied to vectors and matrices are applied elementwise.

# Bibliography

- Alain, G. and Bengio, Y. (2012). What regularized auto-encoders learn from the data generating distribution. Technical Report Arxiv report 1211.4246, Université de Montréal. 302
- Alain, G. and Bengio, Y. (2013). What regularized auto-encoders learn from the data generating distribution. In *ICLR'2013*. also arXiv report 1211.4246. 286, 302, 304
- Alain, G., Bengio, Y., Yao, L., Éric Thibodeau-Laufer, Yosinski, J., and Vincent, P. (2015). GSNs: Generative stochastic networks. arXiv:1503.05571. 288
- Amari, S. (1997). Neural learning in structured parameter spaces - natural Riemannian gradient. In *Advances in Neural Information Processing Systems*, pages 127–133. MIT Press. 116
- Anderson, E. (1935). The Irises of the Gaspé Peninsula. *Bulletin of the American Iris Society*, **59**, 2–5. 14
- Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. Technical report, arXiv:1409.0473. 251
- Bahl, L. R., Brown, P., de Souza, P. V., and Mercer, R. L. (1987). Speech recognition with continuous-parameter hidden Markov models. *Computer, Speech and Language*, **2**, 219–234. 48, 229
- Baldi, P. and Brunak, S. (1998). *Bioinformatics, the Machine Learning Approach*. MIT Press. 231
- Baldi, P. and Sadowski, P. J. (2013). Understanding dropout. In *Advances in Neural Information Processing Systems 26*, pages 2814–2822. 153
- Baldi, P., Brunak, S., Frasconi, P., Soda, G., and Pollastri, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, **15**(11), 937–946. 202
- Barron, A. E. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. on Information Theory*, **39**, 930–945. 126
- Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. Oxford University Press. 290
- Basilevsky, A. (1994). *Statistical Factor Analysis and Related Methods: Theory and Applications*. Wiley. 290
- Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I. J., Bergeron, A., Bouchard, N., and Bengio, Y. (2012). Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop. 57

- Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.*, **37**, 1559–1563. [227](#)
- Baxter, J. (1995). Learning internal representations. In *Proceedings of the 8th International Conference on Computational Learning Theory (COLT'95)*, pages 311–320, Santa Cruz, California. ACM Press. [154](#)
- Becker, S. and Hinton, G. (1992). A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, **355**, 161–163. [335](#)
- Belkin, M. and Niyogi, P. (2002). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS'01*, Cambridge, MA. MIT Press. [322](#)
- Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, **15**(6), 1373–1396. [98](#), [340](#)
- Bengio, S. and Bengio, Y. (2000a). Taking on the curse of dimensionality in joint distributions using neural networks. *IEEE Transactions on Neural Networks, special issue on Data Mining and Knowledge Discovery*, **11**(3), 550–557. [207](#)
- Bengio, Y. (1991). *Artificial Neural Networks and their Application to Sequence Recognition*. Ph.D. thesis, McGill University, (Computer Science), Montreal, Canada. [212](#), [231](#)
- Bengio, Y. (1993). A connectionist approach to speech recognition. *International Journal on Pattern Recognition and Artificial Intelligence*, **7**(4), 647–668. [229](#)
- Bengio, Y. (1999a). Markovian models for sequential data. *Neural Computing Surveys*, **2**, 129–162. [229](#)
- Bengio, Y. (1999b). Markovian models for sequential data. *Neural Computing Surveys*, **2**, 129–162. [231](#)
- Bengio, Y. (2009). *Learning deep architectures for AI*. Now Publishers. [95](#), [128](#)
- Bengio, Y. (2013). Estimating or propagating gradients through stochastic neurons. Technical Report arXiv:1305.2982, Universite de Montreal. [275](#)
- Bengio, Y. and Bengio, S. (2000b). Modeling high-dimensional discrete data with multi-layer neural networks. In *NIPS'99*, pages 400–406. MIT Press. [207](#), [209](#), [210](#)
- Bengio, Y. and Delalleau, O. (2009). Justifying and generalizing contrastive divergence. *Neural Computation*, **21**(6), 1601–1621. [302](#), [363](#), [388](#)
- Bengio, Y. and Frasconi, P. (1996). Input/Output HMMs for sequence processing. *IEEE Transactions on Neural Networks*, **7**(5), 1231–1249. [231](#)
- Bengio, Y. and LeCun, Y. (2007a). Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. [95](#)
- Bengio, Y. and LeCun, Y. (2007b). Scaling learning algorithms towards AI. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, editors, *Large Scale Kernel Machines*. MIT Press. [129](#)
- Bengio, Y. and Monperrus, M. (2005). Non-local manifold tangent learning. In *NIPS'04*, pages 129–136. MIT Press. [97](#), [341](#)

- Bengio, Y., De Mori, R., Flammia, G., and Kompe, R. (1991). Phonetically motivated acoustic parameters for continuous speech recognition using artificial neural networks. In *Proceedings of EuroSpeech'91*. 17
- Bengio, Y., De Mori, R., Flammia, G., and Kompe, R. (1992). Global optimization of a neural network-hidden Markov model hybrid. *IEEE Transactions on Neural Networks*, 3(2), 252–259. 229, 231
- Bengio, Y., Frasconi, P., and Simard, P. (1993). The problem of learning long-term dependencies in recurrent networks. In *IEEE International Conference on Neural Networks*, pages 1183–1195, San Francisco. IEEE Press. (invited paper). 163, 218
- Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Tr. Neural Nets*. 163, 164, 210, 216, 218, 219
- Bengio, Y., LeCun, Y., Nohl, C., and Burges, C. (1995). Lerec: A NN/HMM hybrid for on-line handwriting recognition. *Neural Computation*, 7(6), 1289–1303. 231
- Bengio, Y., Ducharme, R., and Vincent, P. (2001a). A neural probabilistic language model. In *NIPS'00*, pages 932–938. MIT Press. 16
- Bengio, Y., Ducharme, R., and Vincent, P. (2001b). A neural probabilistic language model. In *NIPS'2000*, pages 932–938. 248, 249
- Bengio, Y., Ducharme, R., and Vincent, P. (2001c). A neural probabilistic language model. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *NIPS'2000*, pages 932–938. MIT Press. 343, 344
- Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003a). A neural probabilistic language model. *JMLR*, 3, 1137–1155. 248
- Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003b). A neural probabilistic language model. *Journal of Machine Learning Research*, 3, 1137–1155. 343, 344
- Bengio, Y., Delalleau, O., and Le Roux, N. (2006a). The curse of highly variable functions for local kernel machines. In *NIPS'2005*. 94
- Bengio, Y., Larochelle, H., and Vincent, P. (2006b). Non-local manifold Parzen windows. In *NIPS'2005*. MIT Press. 97, 340
- Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007). Greedy layer-wise training of deep networks. In *NIPS'2006*. 16, 308, 311
- Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *ICML'09*. 117
- Bengio, Y., Léonard, N., and Courville, A. (2013a). Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv:1308.3432. 275
- Bengio, Y., Yao, L., Alain, G., and Vincent, P. (2013b). Generalized denoising auto-encoders as generative models. In *NIPS'2013*. 304, 405, 408
- Bengio, Y., Courville, A., and Vincent, P. (2013c). Representation learning: A review and new perspectives. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 35(8), 1798–1828. 333, 403



- Bengio, Y., Thibodeau-Laufer, E., Alain, G., and Yosinski, J. (2014a). Deep generative stochastic networks trainable by backprop. Technical Report arXiv:1306.1091. [275](#)
- Bengio, Y., Thibodeau-Laufer, E., Alain, G., and Yosinski, J. (2014b). Deep generative stochastic networks trainable by backprop. In *Proceedings of the 30th International Conference on Machine Learning (ICML'14)*. [275](#), [405](#), [407](#), [409](#), [410](#)
- Bennett, C. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, **22**(2), 245–268. [357](#)
- Berglund, M. and Raiko, T. (2013). Stochastic gradient estimate variance in contrastive divergence and persistent contrastive divergence. *CoRR*, **abs/1312.6002**. [364](#)
- Bergstra, J. (2011). *Incorporating Complex Cells into Neural Networks for Pattern Classification*. Ph.D. thesis, Université de Montréal. [285](#)
- Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., and Bengio, Y. (2010). Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*. Oral Presentation. [57](#)
- Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, **24**(3), 179–195. [366](#)
- Bishop, C. M. (1994). Mixture density networks. [113](#)
- Bishop, C. M. (1995). Regularization and complexity control in feed-forward networks. In *Proceedings International Conference on Artificial Neural Networks ICANN'95*, volume 1, page 141–148. [149](#)
- Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1989). Learnability and the vapnik–chervonenkis dimension. *Journal of the ACM*, **36**(4), 929–865. [78](#), [79](#)
- Bordes, A., Glorot, X., Weston, J., and Bengio, Y. (2012). Joint learning of words and meaning representations for open-text semantic parsing. *AISTATS'2012*. [205](#)
- Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, New York, NY, USA. ACM. [13](#), [95](#), [109](#)
- Bottou, L. (1991). *Une approche théorique de l'apprentissage connexioniste; applications à la reconnaissance de la parole*. Ph.D. thesis, Université de Paris XI. [231](#)
- Bottou, L. (2011). From machine learning to machine reasoning. Technical report, arXiv.1102.1808. [204](#), [205](#)
- Bottou, L., Fogelman-Soulié, F., Blanchet, P., and Lienard, J. S. (1990). Speaker independent isolated digit recognition: multilayer perceptrons vs dynamic time warping. *Neural Networks*, **3**, 453–465. [231](#)
- Bottou, L., Bengio, Y., and LeCun, Y. (1997). Global training of document processing systems using graph transformer networks. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR'97)*, pages 490–494, Puerto Rico. IEEE. [223](#), [230](#), [231](#), [232](#), [233](#), [235](#)

- Bourlard, H. and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, **59**, 291–294. 282
- Bourlard, H. and Morgan, N. (1993). *Connectionist Speech Recognition. A Hybrid Approach*, volume 247 of *The Kluwer international series in engineering and computer science*. Kluwer Academic Publishers, Boston. 231
- Bourlard, H. and Wellekens, C. (1990). Links between hidden Markov models and multilayer perceptrons. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12**, 1167–1178. 231
- Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA. 65
- Brady, M. L., Raghavan, R., and Slawny, J. (1989). Back-propagation fails to separate where perceptrons succeed. *IEEE Transactions on Circuits and Systems*, **36**, 665–674. 158
- Brand, M. (2003). Charting a manifold. In *NIPS'2002*, pages 961–968. MIT Press. 98, 340
- Breiman, L. (1994). Bagging predictors. *Machine Learning*, **24**(2), 123–140. 142
- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, **16**(3), 199–231. 5
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA. 95
- Brown, P. (1987). *The Acoustic-Modeling problem in Automatic Speech Recognition*. Ph.D. thesis, Dept. of Computer Science, Carnegie-Mellon University. 229
- Brown, P. F., Pietra, V. J. D., DeSouza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-based  $n$ -gram models of natural language. *Computational Linguistics*, **18**, 467–479. 250
- Bucilua, C., Caruana, R., and Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541. ACM. 236
- Carreira-Perpiñan, M. A. and Hinton, G. E. (2005). On contrastive divergence learning. In R. G. Cowell and Z. Ghahramani, editors, *AISTATS'2005*, pages 33–40. Society for Artificial Intelligence and Statistics. 361, 388
- Caruana, R. (1993). Multitask connectionist learning. In *Proc. 1993 Connectionist Models Summer School*, pages 372–379. 154
- Cauchy, A. (1847). Méthode générale pour la résolution de systèmes d'équations simultanées. In *Compte rendu des séances de l'académie des sciences*, pages 536–538. 58
- Cayton, L. (2005). Algorithms for manifold learning. Technical Report CS2008-0923, UCSD. 98, 336
- Chapelle, O., Weston, J., and Schölkopf, B. (2003). Cluster kernels for semi-supervised learning. In *NIPS'02*, pages 585–592, Cambridge, MA. MIT Press. 322
- Chapelle, O., Schölkopf, B., and Zien, A., editors (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA. 322

- Chellapilla, K., Puri, S., and Simard, P. (2006). High Performance Convolutional Neural Networks for Document Processing. In Guy Lorette, editor, *Tenth International Workshop on Frontiers in Handwriting Recognition*, La Baule (France). Université de Rennes 1, Suvisoft. <http://www.suvisoft.com>. 15, 17
- Chen, S. F. and Goodman, J. T. (1999). An empirical study of smoothing techniques for language modeling. *Computer, Speech and Language*, **13**(4), 359–393. 222, 223
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014)*. 216
- Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. (2014). The loss surface of multilayer networks. arXiv 1412.0233. 311
- Ciresan, D., Meier, U., Masci, J., and Schmidhuber, J. (2012). Multi-column deep neural network for traffic sign classification. *Neural Networks*, **32**, 333–338. 128
- Ciresan, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J. (2010). Deep big simple neural nets for handwritten digit recognition. *Neural Computation*, **22**, 1–14. 15, 17
- Coates, A. and Ng, A. Y. (2011). The importance of encoding versus training with sparse coding and vector quantization. In *ICML’2011*. 17
- Coates, A., Lee, H., and Ng, A. Y. (2011). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*. 243
- Coates, A., Huval, B., Wang, T., Wu, D., Catanzaro, B., and Andrew, N. (2013). Deep learning with cots hpc systems. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 1337–1345. JMLR Workshop and Conference Proceedings. 15, 17
- Collobert, R. (2004). *Large Scale Machine Learning*. Ph.D. thesis, Université de Paris VI, LIP6. 109
- Comon, P. (1994). Independent component analysis - a new concept? *Signal Processing*, **36**, 287–314. 291, 292
- Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, **20**, 273–297. 13, 95
- Coupric, C., Farabet, C., Najman, L., and LeCun, Y. (2013). Indoor semantic segmentation using depth information. In *International Conference on Learning Representations (ICLR2013)*. 128
- Courville, A., Bergstra, J., and Bengio, Y. (2011). Unsupervised models of images by spike-and-slab RBMs. In *ICML’11*. 258, 401
- Courville, A., Desjardins, G., Bergstra, J., and Bengio, Y. (2014). The spike-and-slab RBM and extensions to discrete and sparse data distributions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **36**(9), 1874–1887. 401

- Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory, 2nd Edition*. Wiley-Interscience. 42
- Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, **14**, 1–10. 36
- Crick, F. H. C. and Mitchison, G. (1983). The function of dream sleep. *Nature*, **304**, 111–114. 360
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, **2**, 303–314. 331
- Dauphin, Y. and Bengio, Y. (2013). Stochastic ratio matching of RBMs for sparse high-dimensional inputs. In *NIPS26*. NIPS Foundation. 369
- Dauphin, Y., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *NIPS’2014*. 61, 311
- Davis, A., Rubinstein, M., Wadhwa, N., Mysore, G., Durand, F., and Freeman, W. T. (2014). The visual microphone: Passive recovery of sound from video. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, **33**(4), 79:1–79:10. 241
- de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l’institut Henri Poincaré*, **7**, 1–68. 36
- Delalleau, O. and Bengio, Y. (2011). Shallow vs. deep sum-product networks. In *NIPS*. 127, 331, 332
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*. 14
- Deng, J., Berg, A. C., Li, K., and Fei-Fei, L. (2010). What does classifying more than 10,000 image categories tell us? In *Proceedings of the 11th European Conference on Computer Vision: Part V, ECCV’10*, pages 71–84, Berlin, Heidelberg. Springer-Verlag. 14
- Deng, J., Ding, N., Jia, Y., Frome, A., Murphy, K., Bengio, S., Li, Y., Neven, H., and Adam, H. (2014). Large-scale object classification using label relation graphs. In *ECCV’2014*, pages 48–64. 223
- Desjardins, G. and Bengio, Y. (2008). Empirical evaluation of convolutional RBMs for vision. Technical Report 1327, Département d’Informatique et de Recherche Opérationnelle, Université de Montréal. 402
- Desjardins, G., Courville, A., and Bengio, Y. (2011). On tracking the partition function. In *NIPS’2011*. 358
- Do, T.-M.-T. and Artières, T. (2010). Neural conditional random fields. In *International Conference on Artificial Intelligence and Statistics*, pages 177–184. 223
- Donoho, D. L. and Grimes, C. (2003). Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. Technical Report 2003-08, Dept. Statistics, Stanford University. 98, 340

- Doob, J. (1953). *Stochastic processes*. Wiley: New York. 36
- Doya, K. (1993). Bifurcations of recurrent neural networks in gradient descent learning. *IEEE Transactions on Neural Networks*, **1**, 75–80. 164, 210
- Dugas, C., Bengio, Y., Bélisle, F., and Nadeau, C. (2001). Incorporating second-order functional knowledge for better option pricing. In *NIPS'00*, pages 472–478. MIT Press. 109
- Ebrahimi, S., Pal, C., Bouthillier, X., Froumenty, P., Jean, S., Konda, K. R., Vincent, P., Courville, A., and Bengio, Y. (2013). Combining modality specific deep neural network models for emotion recognition in video. In *Emotion Recognition In The Wild Challenge and Workshop (Emotiw2013)*. 128
- El Hihi, S. and Bengio, Y. (1996). Hierarchical recurrent neural networks for long-term dependencies. In *NIPS 8*. MIT Press. 217, 221, 222
- ElHihi, S. and Bengio, Y. (1996). Hierarchical recurrent neural networks for long-term dependencies. In *NIPS'1995*. 213
- Erhan, D., Bengio, Y., Courville, A., Manzagol, P., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *J. Machine Learning Res.* 309, 311, 312, 313
- Farabet, C., LeCun, Y., Kavukcuoglu, K., Culurciello, E., Martini, B., Akselrod, P., and Talay, S. (2011). Large-scale FPGA-based convolutional networks. In R. Bekkerman, M. Bilenko, and J. Langford, editors, *Scaling up Machine Learning: Parallel and Distributed Approaches*. Cambridge University Press. 298
- Farabet, C., Couprie, C., Najman, L., and LeCun, Y. (2013a). Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 128
- Farabet, C., Couprie, C., Najman, L., and LeCun, Y. (2013b). Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**(8), 1915–1929. 223
- Fei-Fei, L., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**(4), 594–611. 319
- Fischer, A. and Igel, C. (2011). Bounding the bias of contrastive divergence learning. *Neural Computation*, **23**(3), 664–73. 388
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188. 14, 74
- Frasconi, P., Gori, M., and Sperduti, A. (1997). On the efficient classification of data structures by neural networks. In *Proc. Int. Joint Conf. on Artificial Intelligence*. 204, 205
- Frasconi, P., Gori, M., and Sperduti, A. (1998). A general framework for adaptive processing of data structures. *IEEE Transactions on Neural Networks*, **9**(5), 768–786. 205
- Frey, B. J. (1998). *Graphical models for machine learning and digital communication*. MIT Press. 206

- Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, **36**, 193–202. [15](#), [16](#), [17](#)
- Garson, J. (1900). The metric system of identification of criminals, as used in in great britain and ireland. *The Journal of the Anthropological Institute of Great Britain and Ireland*, (2), 177–227. [14](#)
- Girosi, F. (1994). Regularization theory, radial basis functions and networks. In V. Cherkassky, J. Friedman, and H. Wechsler, editors, *From Statistics to Neural Networks*, volume 136 of *NATO ASI Series*, pages 166–187. Springer Berlin Heidelberg. [126](#)
- Glorot, X., Bordes, A., and Bengio, Y. (2011a). Deep sparse rectifier neural networks. In *AISTATS'2011*. [109](#), [297](#)
- Glorot, X., Bordes, A., and Bengio, Y. (2011b). Deep sparse rectifier neural networks. In *JMLR W&CP: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*. [130](#), [297](#)
- Glorot, X., Bordes, A., and Bengio, Y. (2011c). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML'2011*. [297](#), [316](#)
- Gong, S., McKenna, S., and Psarrou, A. (2000). *Dynamic Vision: From Images to Face Recognition*. Imperial College Press. [339](#), [342](#)
- Goodfellow, I., Le, Q., Saxe, A., and Ng, A. (2009). Measuring invariances in deep networks. In *NIPS'2009*, pages 646–654. [285](#), [297](#)
- Goodfellow, I., Koenig, N., Muja, M., Pantofaru, C., Sorokin, A., and Takayama, L. (2010). Help me help you: Interfaces for personal robots. In *Proc. of Human Robot Interaction (HRI)*, Osaka, Japan. ACM Press, ACM Press. [71](#)
- Goodfellow, I., Courville, A., and Bengio, Y. (2012). Large-scale feature learning with spike-and-slab sparse coding. In *ICML'2012*. [293](#)
- Goodfellow, I. J. (2010). Technical report: Multidimensional, downsampled convolution for autoencoders. Technical report, Université de Montréal. [187](#)
- Goodfellow, I. J., Courville, A., and Bengio, Y. (2011). Spike-and-slab sparse coding for unsupervised feature discovery. In *NIPS Workshop on Challenges in Learning Hierarchical Models*. [128](#), [317](#)
- Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y. (2013a). Maxout networks. In S. Dasgupta and D. McAllester, editors, *ICML'13*, pages 1319–1327. [130](#), [152](#), [243](#)
- Goodfellow, I. J., Mirza, M., Courville, A., and Bengio, Y. (2013b). Multi-prediction deep Boltzmann machines. In *NIPS'26*. NIPS Foundation. [367](#), [398](#), [399](#)
- Goodfellow, I. J., Courville, A., and Bengio, Y. (2013c). Scaling up spike-and-slab models for unsupervised feature learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**(8), 1902–1914. [401](#)

- Gori, M. and Tesi, A. (1992). On the problem of local minima in backpropagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-14**(1), 76–86. 158
- Gosset, W. S. (1908). The probable error of a mean. *Biometrika*, **6**(1), 1–25. Originally published under the pseudonym “Student”. 14
- Gouws, S., Bengio, Y., and Corrado, G. (2014). Bilbowa: Fast bilingual distributed representations without word alignments. Technical report, arXiv:1410.2455. 320
- Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks*. Studies in Computational Intelligence. Springer. 202, 215, 216, 223
- Graves, A. (2013). Generating sequences with recurrent neural networks. Technical report, arXiv:1308.0850. 114, 215, 217
- Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, **18**(5), 602–610. 202
- Graves, A. and Schmidhuber, J. (2009). Offline handwriting recognition with multidimensional recurrent neural networks. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *NIPS’2008*, pages 545–552. 202
- Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *ICML’2006*, pages 369–376, Pittsburgh, USA. 223
- Graves, A., Liwicki, M., Bunke, H., Schmidhuber, J., and Fernández, S. (2008). Unconstrained on-line handwriting recognition with recurrent neural networks. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *NIPS’2007*, pages 577–584. 202
- Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *ICASSP’2013*, pages 6645–6649. 202, 215, 216
- Gutmann, M. and Hyvarinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of The Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS’10)*. 370
- Haffner, P., Franzini, M., and Waibel, A. (1991). Integrating time alignment and neural networks for high performance continuous speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 105–108, Toronto. 231
- Håstad, J. (1986). Almost optimal lower bounds for small depth circuits. In *Proceedings of the 18th annual ACM Symposium on Theory of Computing*, pages 6–20, Berkeley, California. ACM Press. 127, 332
- Håstad, J. and Goldmann, M. (1991). On the power of small-depth threshold circuits. *Computational Complexity*, **1**, 113–129. 127, 332
- Henaff, M., Jarrett, K., Kavukcuoglu, K., and LeCun, Y. (2011). Unsupervised learning of sparse features for scalable audio classification. In *ISMIR’11*. 298
- Herault, J. and Ans, B. (1984). Circuits neuronaux à synapses modifiables: Décodage de messages composites par apprentissage non supervisé. *Comptes Rendus de l’Académie des Sciences*, **299(III-13)**, 525–528. 291



- Hinton, G. E. (2000). Training products of experts by minimizing contrastive divergence. Technical Report GCNU TR 2000-004, Gatsby Unit, University College London. 361
- Hinton, G. E. and Roweis, S. (2003). Stochastic neighbor embedding. In *NIPS'2002*. 340
- Hinton, G. E. and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, **313**(5786), 504–507. 287, 308, 309
- Hinton, G. E. and Salakhutdinov, R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, **313**, 504–507. 311
- Hinton, G. E. and Zemel, R. S. (1994). Autoencoders, minimum description length, and Helmholtz free energy. In *NIPS'1993*. 282
- Hinton, G. E., Osindero, S., and Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, **18**, 1527–1554. 16, 17, 308, 309, 311, 389
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. Technical report, arXiv:1207.0580. 139
- Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, T.U. München. 163, 210, 218
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, **9**(8), 1735–1780. 215, 216
- Hochreiter, S., Informatik, F. F., Bengio, Y., Frasconi, P., and Schmidhuber, J. (2000). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In J. Kolen and S. Kremer, editors, *Field Guide to Dynamical Recurrent Networks*. IEEE Press. 216
- Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**, 359–366. 331
- Hsu, F.-H. (2002). *Behind Deep Blue: Building the Computer That Defeated the World Chess Champion*. Princeton University Press, Princeton, NJ, USA. 4
- Huang, F. and Ogata, Y. (2002). Generalized pseudo-likelihood estimates for markov random fields on lattice. *Annals of the Institute of Statistical Mathematics*, **54**(1), 1–18. 366
- Hytyniemi, H. (1996). Turing machines are recurrent neural networks. In *STeP'96*, pages 13–24. 193
- Hyvärinen, A. (1999). Survey on independent component analysis. *Neural Computing Surveys*, **2**, 94–128. 291
- Hyvärinen, A. (2005a). Estimation of non-normalized statistical models using score matching. *J. Machine Learning Res.*, **6**. 301
- Hyvärinen, A. (2005b). Estimation of non-normalized statistical models using score matching. *Journal of Machine Learning Research*, **6**, 695–709. 367
- Hyvärinen, A. (2007a). Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Transactions on Neural Networks*, **18**, 1529–1531. 368



- Hyvärinen, A. (2007b). Some extensions of score matching. *Computational Statistics and Data Analysis*, **51**, 2499–2512. 368
- Hyvärinen, A. and Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, **12**(3), 429–439. 292
- Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. Wiley-Interscience. 291
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixture of local experts. *Neural Computation*, **3**, 79–87. 113
- Jaeger, H. (2003). Adaptive nonlinear system identification with echo state networks. In *Advances in Neural Information Processing Systems 15*. 211
- Jaeger, H. (2007a). Discovering multiscale dynamical features with hierarchical echo state networks. Technical report, Jacobs University. 217
- Jaeger, H. (2007b). Echo state network. *Scholarpedia*, **2**(9), 2330. 210
- Jaeger, H. and Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, **304**(5667), 78–80. 17, 210
- Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., Mooij, J. M., and Schölkopf, B. (2012). On causal and anticausal learning. In *ICML'2012*, pages 1255–1262. 324
- Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009a). What is the best multi-stage architecture for object recognition? In *Proc. International Conference on Computer Vision (ICCV'09)*, pages 2146–2153. IEEE. 15, 17, 129, 130
- Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009b). What is the best multi-stage architecture for object recognition? In *ICCV'09*. 109, 298
- Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, **78**, 2690–2693. 357
- Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press. 35
- Jelinek, F. and Mercer, R. L. (1980). Interpolated estimation of markov source parameters from sparse data. In E. S. Gelsema and L. N. Kanal, editors, *Pattern Recognition in Practice*. North-Holland, Amsterdam. 222
- Jordan, M. I. (1998). *Learning in Graphical Models*. Kluwer, Dordrecht, Netherlands. 13
- Juang, B. H. and Katagiri, S. (1992). Discriminative learning for minimum error classification. *IEEE Transactions on Signal Processing*, **40**(12), 3043–3054. 229
- Jutten, C. and Herault, J. (1991). Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture. *Signal Processing*, **24**, 1–10. 291
- Kamyschanska, H. and Memisevic, R. (2015). The potential energy of an autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 304
- Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *CVPR*. 14

- Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **ASSP-35**(3), 400–401. 222
- Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2008a). Fast inference in sparse coding algorithms with applications to object recognition. CBLL-TR-2008-12-01, NYU. 285
- Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2008b). Fast inference in sparse coding algorithms with applications to object recognition. Technical report, Computational and Biological Learning Lab, Courant Institute, NYU. Tech Report CBLL-TR-2008-12-01. 298
- Kavukcuoglu, K., Ranzato, M.-A., Fergus, R., and LeCun, Y. (2009). Learning invariant features through topographic filter maps. In *CVPR'2009*. 298
- Kavukcuoglu, K., Sermanet, P., Boureau, Y.-L., Gregor, K., Mathieu, M., and LeCun, Y. (2010). Learning convolutional feature hierarchies for visual recognition. In *NIPS'2010*. 298
- Kindermann, R. (1980). *Markov Random Fields and Their Applications (Contemporary Mathematics ; V. 1)*. American Mathematical Society. 261
- Kingma, D. and LeCun, Y. (2010a). Regularized estimation of image statistics by score matching. In *NIPS'2010*. 301
- Kingma, D. and LeCun, Y. (2010b). Regularized estimation of image statistics by score matching. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1126–1134. 369
- Kingma, D., Rezende, D., Mohamed, S., and Welling, M. (2014). Semi-supervised learning with deep generative models. In *NIPS'2014*. 275
- Kingma, D. P. (2013). Fast gradient-based inference with continuous latent variable models in auxiliary form. Technical report, arxiv:1306.0733. 275
- Kingma, D. P. and Welling, M. (2014a). Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 275, 342, 343
- Kingma, D. P. and Welling, M. (2014b). Efficient gradient-based inference through transformations between bayes nets and neural nets. Technical report, arxiv:1402.0480. 275
- Klementiev, A., Titov, I., and Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*. 320
- Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press. 227, 273, 279
- Koren, Y. (2009). 1 the bellkor solution to the netflix grand prize. 143
- Koutnik, J., Greff, K., Gomez, F., and Schmidhuber, J. (2014). A clockwork RNN. In *ICML'2014*. 217, 222
- Krause, O., Fischer, A., Glasmachers, T., and Igel, C. (2013). Approximation properties of DBNs with binary hidden units and real-valued visible units. In *ICML'2013*. 331
- Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto. 14, 258

- Krizhevsky, A., Sutskever, I., and Hinton, G. (2012a). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (NIPS'2012)*. 15, 17, 71
- Krizhevsky, A., Sutskever, I., and Hinton, G. (2012b). ImageNet classification with deep convolutional neural networks. In *NIPS'2012*. 128, 297
- Lafferty, J., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In C. E. Brodley and A. P. Danyluk, editors, *ICML 2001*. Morgan Kaufmann. 223, 229
- Lang, K. J. and Hinton, G. E. (1988). The development of the time-delay neural network architecture for speech recognition. Technical Report CMU-CS-88-152, Carnegie-Mellon University. 191, 212
- Lappalainen, H., Giannakopoulos, X., Honkela, A., and Karhunen, J. (2000). Nonlinear independent component analysis using ensemble learning: Experiments and discussion. In *Proc. ICA*. Citeseer. 292
- Larochelle, H. and Bengio, Y. (2008a). Classification using discriminative restricted Boltzmann machines. In *ICML'2008*. 285, 411
- Larochelle, H. and Bengio, Y. (2008b). Classification using discriminative restricted Boltzmann machines. In *ICML'08*, pages 536–543. ACM. 322
- Larochelle, H. and Murray, I. (2011). The Neural Autoregressive Distribution Estimator. In *AISTATS'2011*. 205, 209
- Larochelle, H., Erhan, D., and Bengio, Y. (2008). Zero-data learning of new tasks. In *AAAI Conference on Artificial Intelligence*. 319
- Lasserre, J. A., Bishop, C. M., and Minka, T. P. (2006). Principled hybrids of generative and discriminative models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR'06)*, pages 87–94, Washington, DC, USA. IEEE Computer Society. 322
- Le, Q., Ranzato, M., Monga, R., Devin, M., Corrado, G., Chen, K., Dean, J., and Ng, A. (2012). Building high-level features using large scale unsupervised learning. In *ICML'2012*. 15, 17
- Le Roux, N. and Bengio, Y. (2010). Deep belief networks are compact universal approximators. *Neural Computation*, 22(8), 2192–2207. 331
- Le Roux, N., Manzagol, P.-A., and Bengio, Y. (2008). Topmoumoute online natural gradient algorithm. In *NIPS'07*. 116
- LeCun, Y. (1987). *Modèles connexionistes de l'apprentissage*. Ph.D. thesis, Université de Paris VI. 13, 282
- LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4), 541–551. 16
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998a). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. 13, 14, 223, 230, 232

- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998b). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**(11), 2278–2324. [17](#)
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998c). Gradient based learning applied to document recognition. *Proc. IEEE*. [16](#)
- Lee, H., Ekanadham, C., and Ng, A. (2008). Sparse deep belief net model for visual area V2. In *NIPS'07*. [285](#)
- Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In L. Bottou and M. Littman, editors, *ICML 2009*. ACM, Montreal, Canada. [402](#)
- Lenat, D. B. and Guha, R. V. (1989). *Building large knowledge-based systems; representation and inference in the Cyc project*. Addison-Wesley Longman Publishing Co., Inc. [5](#)
- Leprieur, H. and Haffner, P. (1995). Discriminant learning with minimum memory loss for improved non-vocabulary rejection. In *EUROSPEECH'95*, Madrid, Spain. [229](#)
- Lin, T., Horne, B. G., Tino, P., and Giles, C. L. (1996). Learning long-term dependencies is not as difficult with NARX recurrent neural networks. *IEEE Transactions on Neural Networks*, **7**(6), 1329–1338. [213](#)
- Linde, N. (1992). The machine that changed the world, episode 3. Documentary miniseries. [5](#)
- Long, P. M. and Servedio, R. A. (2010). Restricted Boltzmann machines are hard to approximately evaluate or simulate. In *Proceedings of the 27th International Conference on Machine Learning (ICML'10)*. [384](#)
- Lovelace, A. (1842). Notes upon L. F. Menabrea's "Sketch of the Analytical Engine invented by Charles Babbage". [4](#)
- Lowerre, B. (1976). *The Harpy Speech Recognition System*. Ph.D. thesis. [224](#), [229](#), [233](#)
- Lukoševičius, M. and Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, **3**(3), 127–149. [210](#)
- Luo, H., Carrier, P.-L., Courville, A., and Bengio, Y. (2013). Texture modeling with convolutional spike-and-slab RBMs and deep extensions. In *AISTATS'2013*. [72](#)
- Lyu, S. (2009). Interpretation and generalization of score matching. In *UAI'09*. [368](#)
- Maass, W., Natschlaeger, T., and Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, **14**(11), 2531–2560. [210](#)
- MacKay, D. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press. [42](#)
- Marlin, B., Swersky, K., Chen, B., and de Freitas, N. (2010). Inductive principles for restricted Boltzmann machine learning. In *Proceedings of The Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS'10)*, volume 9, pages 509–516. [364](#), [368](#), [369](#), [385](#)

- Martens, J. and Medabalimi, V. (2014). On the expressive efficiency of sum product networks. *arXiv:1411.7717*. 332
- Martens, J. and Sutskever, I. (2011). Learning recurrent neural networks with Hessian-free optimization. In *Proc. ICML'2011*. ACM. 219
- Mase, S. (1995). Consistency of the maximum pseudo-likelihood estimator of continuous state space Gibbsian processes. *The Annals of Applied Probability*, 5(3), pp. 603–612. 366
- Matan, O., Burges, C. J. C., LeCun, Y., and Denker, J. S. (1992). Multi-digit recognition using a space displacement neural network. In *NIPS'91*, pages 488–495, San Mateo CA. Morgan Kaufmann. 231
- McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London. 110
- Mesnil, G., Dauphin, Y., Glorot, X., Rifai, S., Bengio, Y., Goodfellow, I., Lavoie, E., Muller, X., Desjardins, G., Warde-Farley, D., Vincent, P., Courville, A., and Bergstra, J. (2011). Unsupervised and transfer learning challenge: a deep learning approach. In *JMLR W&CP: Proc. Unsupervised and Transfer Learning*, volume 7. 128, 317
- Mesnil, G., Rifai, S., Dauphin, Y., Bengio, Y., and Vincent, P. (2012). Surfing on the manifold. Learning Workshop, Snowbird. 404
- Mikolov, T. (2012). *Statistical Language Models based on Neural Networks*. Ph.D. thesis, Brno University of Technology. 114, 220
- Mikolov, T., Le, Q. V., and Sutskever, I. (2013). Exploiting similarities among languages for machine translation. Technical report, arXiv:1309.4168. 320
- Minka, T. (2005). Divergence measures and message passing. *Microsoft Research Cambridge UK Tech Rep MSRTR2005173*, 72(TR-2005-173). 354
- Minsky, M. L. and Papert, S. A. (1969). *Perceptrons*. MIT Press, Cambridge. 13
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York. 70
- Mnih, A. and Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2265–2273. Curran Associates, Inc. 371
- Montúfar, G. (2014). Universal approximation depth and errors of narrow belief networks with discrete units. *Neural Computation*, 26. 331
- Montúfar, G. and Ay, N. (2011). Refinements of universal approximation results for deep belief networks and restricted Boltzmann machines. *Neural Computation*, 23(5), 1306–1319. 331
- Montufar, G. and Morton, J. (2014). When does a mixture of products contain a product of mixtures? *SIAM Journal on Discrete Mathematics (SIDMA)*. 330
- Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. (2014). On the number of linear regions of deep neural networks. In *NIPS'2014*. 329, 332, 333

- Mor-Yosef, S., Samueloff, A., Modan, B., Navot, D., and Schenker, J. G. (1990). Ranking the risk factors for cesarean: logistic regression analysis of a nationwide study. *Obstet Gynecol*, **75**(6), 944–7. [5](#)
- Mozer, M. C. (1992). The induction of multiscale temporal structure. In *NIPS'91*, pages 275–282, San Mateo, CA. Morgan Kaufmann. [213](#), [222](#)
- Murphy, K. P. (2012). *Machine Learning: a Probabilistic Perspective*. MIT Press, Cambridge, MA, USA. [111](#)
- Murray, B. U. I. and Larochelle, H. (2014). A deep and tractable density estimator. In *ICML'2014*. [114](#), [209](#), [210](#)
- Nadas, A., Nahamoo, D., and Picheny, M. A. (1988). On a model-robust training method for speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-36**(9), 1432–1436. [229](#)
- Nair, V. and Hinton, G. (2010). Rectified linear units improve restricted Boltzmann machines. In *ICML'2010*. [109](#), [297](#)
- Narayanan, H. and Mitter, S. (2010). Sample complexity of testing the manifold hypothesis. In *NIPS'2010*. [98](#), [336](#)
- Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics. Springer. [153](#)
- Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, **11**(2), 125–139. [356](#), [357](#)
- Neal, R. M. (2005). Estimating ratios of normalizing constants using linked importance sampling. [357](#), [358](#)
- Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. (2011). Reading digits in natural images with unsupervised feature learning. Deep Learning and Unsupervised Feature Learning Workshop, NIPS. [14](#)
- Ney, H. and Kneser, R. (1993). Improved clustering techniques for class-based statistical language modelling. In *European Conference on Speech Communication and Technology (Eurospeech)*, pages 973–976, Berlin. [250](#)
- Niesler, T. R., Whittaker, E. W. D., and Woodland, P. C. (1998). Comparison of part-of-speech and automatically derived category-based language models for speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 177–180. [250](#)
- Niranjan, M. and Fallside, F. (1990). Neural networks and radial basis functions in classifying static speech patterns. *Computer Speech and Language*, **4**, 275–289. [109](#)
- Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer. [65](#), [68](#)
- Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, **381**, 607–609. [285](#), [335](#)

- Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Research*, **37**, 3311–3325. 296
- Park, H., Amari, S.-I., and Fukumizu, K. (2000). Adaptive natural gradient learning algorithms for various stochastic models. *Neural Networks*, **13**(7), 755 – 764. 116
- Pascanu, R. (2014). *On recurrent and deep networks*. Ph.D. thesis, Université de Montréal. 160, 161
- Pascanu, R. and Bengio, Y. (2012). On the difficulty of training recurrent neural networks. Technical Report arXiv:1211.5063, Université de Montréal. 114
- Pascanu, R. and Bengio, Y. (2013). Revisiting natural gradient for deep networks. Technical report, arXiv:1301.3584. 116
- Pascanu, R., Mikolov, T., and Bengio, Y. (2013a). On the difficulty of training recurrent neural networks. In *ICML’2013*. 114, 164, 210, 213, 220, 221, 222
- Pascanu, R., Montufar, G., and Bengio, Y. (2013b). On the number of inference regions of deep feed forward networks with piece-wise linear activations. Technical report, U. Montreal, arXiv:1312.6098. 127
- Pascanu, R., Güleşhre, Ç., Cho, K., and Bengio, Y. (2014a). How to construct deep recurrent neural networks. In *ICLR’2014*. 153
- Pascanu, R., Gulcehre, C., Cho, K., and Bengio, Y. (2014b). How to construct deep recurrent neural networks. In *ICLR’2014*. 215, 217, 332
- Pascanu, R., Montufar, G., and Bengio, Y. (2014c). On the number of inference regions of deep feed forward networks with piece-wise linear activations. In *ICLR’2014*. 329
- Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine*, pages 329–334. 259
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann. 36
- Petersen, K. B. and Pedersen, M. S. (2006). The matrix cookbook. Version 20051003. 20
- Pinto, N., Cox, D. D., and DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Comput Biol*, **4**. 402
- Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, **46**(1), 77–105. 204
- Polyak, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, **4**(5), 1–17. 166
- Poon, H. and Domingos, P. (2011). Sum-product networks: A new deep architecture. In *UAI’2011*, Barcelona, Spain. 127, 331, 332
- Poundstone, W. (2005). *Fortune’s Formula: The untold story of the scientific betting system that beat the casinos and Wall Street*. Macmillan. 42



- Powell, M. (1987). Radial basis functions for multivariable interpolation: A review. [109](#)
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**(2), 257–286. [227](#)
- Rabiner, L. R. and Juang, B. H. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, pages 257–285. [191](#), [227](#)
- Raiko, T., Yao, L., Cho, K., and Bengio, Y. (2014). Iterative neural autoregressive distribution estimator (NADE-k). Technical report, arXiv:1406.1485. [209](#)
- Raina, R., Madhavan, A., and Ng, A. Y. (2009). Large-scale deep unsupervised learning using graphics processors. In L. Bottou and M. Littman, editors, *ICML 2009*, pages 873–880, New York, NY, USA. ACM. [17](#)
- Ramsey, F. P. (1926). Truth and probability. In R. B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*, chapter 7, pages 156–198. McMaster University Archive for the History of Economic Thought. [37](#)
- Ranzato, M., Poultney, C., Chopra, S., and LeCun, Y. (2007). Efficient learning of sparse representations with an energy-based model. In *NIPS’2006*. [16](#), [296](#), [308](#), [309](#), [311](#)
- Ranzato, M., Boureau, Y., and LeCun, Y. (2008). Sparse feature learning for deep belief networks. In *NIPS’2007*. [296](#)
- Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *ICML’2014*. [275](#)
- Richard Socher, Milind Ganjoo, C. D. M. and Ng, A. Y. (2013). Zero-shot learning through cross-modal transfer. In *27th Annual Conference on Neural Information Processing Systems (NIPS 2013)*. [319](#), [320](#)
- Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. (2011a). Contractive auto-encoders: Explicit invariance during feature extraction. In *ICML’2011*. [304](#), [305](#), [306](#), [338](#)
- Rifai, S., Mesnil, G., Vincent, P., Muller, X., Bengio, Y., Dauphin, Y., and Glorot, X. (2011b). Higher order contractive auto-encoder. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*. [285](#)
- Rifai, S., Mesnil, G., Vincent, P., Muller, X., Bengio, Y., Dauphin, Y., and Glorot, X. (2011c). Higher order contractive auto-encoder. In *ECML PKDD*. [304](#)
- Rifai, S., Dauphin, Y., Vincent, P., Bengio, Y., and Muller, X. (2011d). The manifold tangent classifier. In *NIPS’2011*. [350](#), [351](#)
- Rifai, S., Bengio, Y., Dauphin, Y., and Vincent, P. (2012). A generative process for sampling contractive auto-encoders. In *ICML’2012*. [404](#)
- Roberts, S. and Everson, R. (2001). *Independent component analysis: principles and practice*. Cambridge University Press. [292](#)
- Robinson, A. J. and Fallside, F. (1991). A recurrent error propagation network speech recognition system. *Computer Speech and Language*, **5**(3), 259–274. [17](#)



- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**, 386–408. [13](#), [17](#)
- Rosenblatt, F. (1962). *Principles of Neurodynamics*. Spartan, New York. [13](#), [17](#)
- Roweis, S. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**(5500). [98](#), [340](#)
- Rumelhart, D., Hinton, G., and Williams, R. (1986a). Learning representations by back-propagating errors. *Nature*, **323**, 533–536. [13](#), [248](#)
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986b). Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, volume 1, chapter 8, pages 318–362. MIT Press, Cambridge. [14](#), [17](#)
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986c). Learning representations by back-propagating errors. *Nature*, **323**, 533–536. [104](#), [191](#)
- Rumelhart, D. E., McClelland, J. L., and the PDP Research Group (1986d). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1. MIT Press, Cambridge. [104](#)
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2014). ImageNet Large Scale Visual Recognition Challenge. [14](#)
- Salakhutdinov, R. and Hinton, G. (2009a). Deep Boltzmann machines. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 5, pages 448–455. [15](#), [17](#), [309](#), [391](#), [395](#), [398](#)
- Salakhutdinov, R. and Hinton, G. (2009b). Deep Boltzmann machines. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS 2009)*, volume 8. [400](#), [409](#)
- Salakhutdinov, R. and Hinton, G. E. (2008). Using deep belief nets to learn covariance kernels for Gaussian processes. In *NIPS’07*, pages 1249–1256, Cambridge, MA. MIT Press. [322](#)
- Salakhutdinov, R. and Murray, I. (2008). On the quantitative analysis of deep belief networks. In W. W. Cohen, A. McCallum, and S. T. Roweis, editors, *ICML 2008*, volume 25, pages 872–879. ACM. [357](#)
- Saul, L. K., Jaakkola, T., and Jordan, M. I. (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, **4**, 61–76. [17](#)
- Schaul, T., Zhang, S., and LeCun, Y. (2012). No More Pesky Learning Rates. Technical report, New York University, arxiv 1206.1106. [171](#)
- Schmidhuber, J. (1992). Learning complex, extended sequences using the principle of history compression. *Neural Computation*, **4**(2), 234–242. [16](#), [217](#)
- Schölkopf, B. and Smola, A. (2002). *Learning with kernels*. MIT Press. [95](#)
- Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, **10**, 1299–1319. [98](#), [340](#)

- Schölkopf, B., Burges, C. J. C., and Smola, A. J. (1999). *Advances in Kernel Methods — Support Vector Learning*. MIT Press, Cambridge, MA. 13, 109, 128
- Schulz, H. and Behnke, S. (2012). Learning two-layer contractive encodings. In *ICANN'2012*, pages 620–628. 305
- Schuster, M. and Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673–2681. 202
- Sermanet, P., Kavukcuoglu, K., Chintala, S., and LeCun, Y. (2013). Pedestrian detection with unsupervised multi-stage feature learning. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR'13)*. IEEE. 128
- Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2014). Overfeat: Integrated recognition, localization and detection using convolutional networks. *International Conference on Learning Representations*. 71
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423. 42
- Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the Institute of Radio Engineers*, 37(1), 10–21. 42
- Shilov, G. (1977). *Linear Algebra*. Dover Books on Mathematics Series. Dover Publications. 20
- Siegelmann, H. (1995). Computation beyond the Turing limit. *Science*, 268(5210), 545–548. 193
- Siegelmann, H. and Sontag, E. (1991). Turing computability with neural nets. *Applied Mathematics Letters*, 4(6), 77–80. 193
- Siegelmann, H. T. and Sontag, E. D. (1995). On the computational power of neural nets. *Journal of Computer and Systems Sciences*, 50(1), 132–150. 164
- Simard, P., Victorri, B., LeCun, Y., and Denker, J. (1992). Tangent prop - A formalism for specifying selected invariances in an adaptive network. In *NIPS'1991*. 349, 350, 351
- Simard, P. Y., LeCun, Y., and Denker, J. (1993). Efficient pattern recognition using a new transformation distance. In *NIPS'92*. 348
- Simard, P. Y., LeCun, Y. A., Denker, J. S., and Victorri, B. (1998). Transformation invariance in pattern recognition — tangent distance and tangent propagation. *Lecture Notes in Computer Science*, 1524. 348
- Sjöberg, J. and Ljung, L. (1995). Overtraining, regularization and searching for a minimum, with application to neural networks. *International Journal of Control*, 62(6), 1391–1407. 149
- Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, volume 1, chapter 6, pages 194–281. MIT Press, Cambridge. 266, 277
- Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., and Manning, C. D. (2011a). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *NIPS'2011*. 205

- Socher, R., Manning, C., and Ng, A. Y. (2011b). Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the Twenty-Eighth International Conference on Machine Learning (ICML'2011)*. 205
- Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. (2011c). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *EMNLP'2011*. 205
- Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP'2013*. 205
- Solla, S. A., Levin, E., and Fleisher, M. (1988). Accelerated learning in layered neural networks. *Complex Systems*, **2**, 625–639. 112
- Sontag, E. D. and Sussman, H. J. (1989). Backpropagation can give rise to spurious local minima even for networks without hidden layers. *Complex Systems*, **3**, 91–106. 158
- Srivastava, N. and Salakhutdinov, R. (2012). Multimodal learning with deep Boltzmann machines. In *NIPS'2012*. 321
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, **15**, 1929–1958. 151, 152, 153, 398
- Stewart, L., He, X., and Zemel, R. S. (2007). Learning flexible features for conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**(8), 1415–1426. 223
- Sutskever, I. (2012). *Training Recurrent Neural Networks*. Ph.D. thesis, Departement of computer science, University of Toronto. 211, 219
- Sutskever, I. and Tieleman, T. (2010). On the Convergence Properties of Contrastive Divergence. In Y. W. Teh and M. Titterton, editors, *Proc. of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9, pages 789–795. 364
- Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *ICML*. 167, 211, 219
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. Technical report, arXiv:1409.3215. 215, 216
- Swersky, K. (2010). *Inductive Principles for Learning Restricted Boltzmann Machines*. Master's thesis, University of British Columbia. 302
- Swersky, K., Ranzato, M., Buchman, D., Marlin, B., and de Freitas, N. (2011). On autoencoders and score matching for energy based models. In *ICML'2011*. ACM. 369
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going deeper with convolutions. Technical report, arXiv:1409.4842. 15, 17
- Tenenbaum, J., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**(5500), 2319–2323. 98, 312, 313, 340

- Thrun, S. (1995). Learning to play the game of chess. In *NIPS'1994*. 350
- Tibshirani, R. J. (1995). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **58**, 267–288. 137
- Tieleman, T. (2008). Training restricted Boltzmann machines using approximations to the likelihood gradient. In W. W. Cohen, A. McCallum, and S. T. Roweis, editors, *ICML 2008*, pages 1064–1071. ACM. 364, 388
- Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal components analysis. *Journal of the Royal Statistical Society B*, **61**(3), 611–622. 291
- Uria, B., Murray, I., and Larochelle, H. (2013). Rnade: The real-valued neural autoregressive density-estimator. In *NIPS'2013*. 208, 209
- Utgoff, P. E. and Stracuzzi, D. J. (2002). Many-layered learning. *Neural Computation*, **14**, 2497–2539. 16
- van der Maaten, L. and Hinton, G. E. (2008a). Visualizing data using t-SNE. *J. Machine Learning Res.*, **9**. 312, 340, 343
- van der Maaten, L. and Hinton, G. E. (2008b). Visualizing data using t-SNE. *Journal of Machine Learning Research*, **9**, 2579–2605. 313
- Vapnik, V. N. (1982). *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, Berlin. 78, 79
- Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York. 78, 79, 81
- Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, **16**, 264–280. 78, 79
- Vincent, P. (2011a). A connection between score matching and denoising autoencoders. *Neural Computation*, **23**(7). 301, 302, 304, 404
- Vincent, P. (2011b). A connection between score matching and denoising autoencoders. *Neural Computation*, **23**(7), 1661–1674. 369, 405
- Vincent, P. and Bengio, Y. (2003). Manifold Parzen windows. In *NIPS'2002*. MIT Press. 340
- Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *ICML 2008*. 298
- Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Machine Learning Res.*, **11**. 298
- Wager, S., Wang, S., and Liang, P. (2013). Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems 26*, pages 351–359. 153
- Waibel, A., Hanazawa, T., Hinton, G. E., Shikano, K., and Lang, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**, 328–339. 191

- Wan, L., Zeiler, M., Zhang, S., LeCun, Y., and Fergus, R. (2013). Regularization of neural networks using dropconnect. In *ICML'2013*. 154
- Wang, S. and Manning, C. (2013). Fast dropout training. In *ICML'2013*. 153
- Warde-Farley, D., Goodfellow, I. J., Courville, A., and Bengio, Y. (2014). An empirical analysis of dropout in piecewise linear networks. In *ICLR'2014*. 153
- Weinberger, K. Q. and Saul, L. K. (2004). Unsupervised learning of image manifolds by semidefinite programming. In *CVPR'2004*, pages 988–995. 98, 340
- Weston, J., Ratle, F., and Collobert, R. (2008). Deep learning via semi-supervised embedding. In W. W. Cohen, A. McCallum, and S. T. Roweis, editors, *ICML 2008*, pages 1168–1175, New York, NY, USA. ACM. 322
- Weston, J., Bengio, S., and Usunier, N. (2010). Large scale image annotation: learning to rank with joint word-image embeddings. *Machine Learning*, 81(1), 21–35. 205
- White, H. (1990). Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, 3(5), 535–549. 126
- Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. In *1960 IRE WESCON Convention Record*, volume 4, pages 96–104. IRE, New York. 13, 14, 15, 17
- Wikipedia (2015). List of animals by number of neurons — wikipedia, the free encyclopedia. [Online; accessed 4-March-2015]. 15, 17
- Williams, C. K. I. and Rasmussen, C. E. (1996). Gaussian processes for regression. In *NIPS'95*, pages 514–520. MIT Press, Cambridge, MA. 128
- Wolpert, D. H. (1996). The lack of a priori distinction between learning algorithms. *Neural Computation*, 8(7), 1341–1390. 127
- Xiong, H. Y., Barash, Y., and Frey, B. J. (2011). Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics*, 27(18), 2554–2562. 153
- Xu, L. and Jordan, M. I. (1996). On convergence properties of the EM algorithm for gaussian mixtures. *Neural Computation*, 8, 129–151. 228
- Younes, L. (1998). On the convergence of Markovian stochastic algorithms with rapidly decreasing ergodicity rates. In *Stochastics and Stochastics Models*, pages 177–228. 364, 388
- Zaslavsky, T. (1975). *Facing Up to Arrangements: Face-Count Formulas for Partitions of Space by Hyperplanes*. Number no. 154 in *Memoirs of the American Mathematical Society*. American Mathematical Society. 330
- Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *ECCV'14*. 9, 71
- Zhou, J. and Troyanskaya, O. G. (2014). Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. In *ICML'2014*. 410
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2), 301–320. 116
- Zöhrer, M. and Pernkopf, F. (2014). General stochastic networks for classification. In *NIPS'2014*. 411

# Index

- $L^p$  norm, 26
- Active constraint, 68
- ADALINE, *see* Adaptive Linear Element
- Adaptive Linear Element, 13, 15, 17
- AIS, *see* annealed importance sampling
- Almost everywhere, 52
- Ancestral sampling, 279
- Annealed importance sampling, 354, 396
- Approximate inference, 274
- Artificial intelligence, 4
- Asymptotically unbiased, 84
- Autoencoder, 7
- Bagging, 142
- Bayes' rule, 51
- Bayesian network, *see* directed graphical model
- Bayesian probability, 37
- Beam Search, 233
- Belief network, *see* directed graphical model
- Bernoulli distribution, 44
- Boltzmann distribution, 265
- Boltzmann machine, 265
- Broadcasting, 22
- Calculus of variations, 378
- CD, *see* contrastive divergence
- Centering trick (DBM), 398
- Central limit theorem, 45
- Chain rule of probability, 40
- Chess, 4
- Chord, 272
- Chordal graph, 272
- Classical regularization, 132
- Classification, 71
- Cliffs, 159
- Clipping the gradient, 220
- Clique potential, *see* factor (graphical model)
- CNN, *see* convolutional neural network
- Collider, *see* explaining away
- Computer vision, 241
- Conditional computation, *see* dynamically structured nets, 237
- Conditional independence, 40
- Conditional probability, 39
- Constrained optimization, 67
- Context-specific independence, 267
- Contrast, 242
- Contrastive divergence, 361, 397, 398
- Convolution, 173, 401
- Convolutional neural network, 173
- Coordinate descent, 168, 169, 398
- Correlation, 41
- Cost function, *see* objective function
- Covariance, 41
- Covariance matrix, 42
- curse of dimensionality, 98
- Cyc, 5
- D-separation, 266
- Dataset augmentation, 242, 247
- DBM, *see* deep Boltzmann machine
- Decoder, 7
- Deep belief network, 17, 372, 383, 389, 402
- Deep Blue, 4
- Deep Boltzmann machine, 15, 17, 372, 383, 391, 398, 402
- Deep learning, 4, 7
- Denoising score matching, 369
- Density estimation, 71
- Derivative, 58
- Detector layer, 178
- Dirac delta function, 47
- Directed graphical model, 259
- Directional derivative, 62
- Distributed Representation, 325
- domain adaptation, 315
- Dot product, 23

- Doubly block circulant matrix, 175
- Dream sleep, 360, 381
- DropConnect, 154
- Dropout, 151, 398
- Dynamically structured networks, 237
  
- E-step, 375
- Early stopping, 116, 143, 146, 148
- EBM, *see* energy-based model
- Echo state network, 15, 17
- Effective number of parameters, 134
- Eigendecomposition, 29
- Eigenvalue, 29
- Eigenvector, 29
- ELBO, *see* evidence lower bound
- Element-wise product, *see* Hadamard product
- EM, *see* expectation maximization
- Embedding, 339
- Empirical distribution, 47
- Empirical risk, 156
- Empirical risk minimization, 157
- Encoder, 7
- Energy function, 265
- Energy-based model, 265, 392
- Ensemble methods, 142
- Epoch, 158, 166
- Equality constraint, 67
- Equivariance, 176
- Error function, *see* objective function
- Euclidean norm, 27
- Euler-Lagrange equation, 379
- Evidence lower bound, 372, 374–376, 391
- Expectation, 41
- Expectation maximization, 375
- Expected value, *see* expectation
- Explaining away, 268
  
- Factor (graphical model), 262
- Factor graph, 272
- Factors of variation, 7
- Frequentist probability, 37
- Functional derivatives, 378
  
- Gaussian distribution, *see* Normal distribution 45
- Gaussian mixture, 48
- GCN, *see* Global contrast normalization
- Generalized Lagrange function, *see* Generalized Lagrangian
- Generalized Lagrangian, 67
  
- Gibbs distribution, 263
- Gibbs sampling, 280
- Global contrast normalization, 243
- Gradient, 62
- Gradient clipping, 220
- Gradient descent, 62
- Graph Transformer, 232
- Graphical model, *see* structured probabilistic model
- Greedy layer-wise unsupervised pre-training, 308
  
- Hadamard product, 23
- Harmonium, *see* Restricted Boltzmann machine 277
- Harmony theory, 266
- Helmholtz free energy, *see* evidence lower bound
- Hessian matrix, 63
- Hidden layer, 9
  
- Identity matrix, 24
- Immortality, 270
- Independence, 40
- Inequality constraint, 67
- Inference, 257, 274, 372, 374–376, 378, 380
- Invariance, 181
  
- Jacobian matrix, 52, 62
- Joint probability, 38
  
- Karush-Kuhn-Tucker conditions, 68
- Karush-Kuhn-Tucker, 67
- Kernel (convolution), 174
- KKT, *see* Karush-Kuhn-Tucker
- KKT conditions, *see* Karush-Kuhn-Tucker conditions
- KL divergence, *see* Kullback-Leibler divergence 43
- Knowledge base, 5
- Kullback-Leibler divergence, 43
  
- Lagrange multipliers, 67, 68, 379
- Lagrangian, *see* Generalized Lagrangian 67
- Latent variable, 286
- Line search, 62
- Linear combination, 25
- Linear dependence, 26
- Local conditional probability distribution, 260
- Logistic regression, 5
- Logistic sigmoid, 48
- Loop, 272
- Loss function, *see* objective function



- M-step, 375
- Machine learning, 5
- Manifold hypothesis, 336
- manifold hypothesis, 98
- Manifold learning, 97, 336
- MAP inference, 376
- Marginal probability, 39
- Markov chain, 279
- Markov network, *see* undirected model261
- Markov random field, *see* undirected model261
- Matrix, 21
- Matrix inverse, 24
- Matrix product, 22
- Max pooling, 181
- Mean field, 397, 398
- Measure theory, 51
- Measure zero, 52
- Method of steepest descent, *see* gradient descent
- Missing inputs, 71
- Mixing (Markov chain), 281
- Mixture distribution, 48
- MLP, *see* multilayer perception
- MNIST, 398
- Model averaging, 142
- Model compression, 236
- Moore-Penrose pseudoinverse, 140
- Moralized graph, 270
- MP-DBM, *see* multi-prediction DBM
- MRF (Markov Random Field), *see* undirected model261
- Multi-modal learning, 321
- Multi-prediction DBM, 397, 398
- Multi-task learning, 154
- Multilayer perception, 8
- Multilayer perceptron, 17
- Multinomial distribution, 44
- Multinoulli distribution, 44
  
- Naive Bayes, 5, 53
- Nat, 42
- natural image, 256
- Negative definite, 63
- Negative phase, 360
- Neocognitron, 15, 17
- Nesterov momentum, 167
- Netflix Grand Prize, 143
- Noise-contrastive estimation, 370
- Norm, 26
- Normal distribution, 45
  
- Normal equations, 135
  
- Object detection, 241
- Object recognition, 241
- Objective function, 58
- one-shot learning, 319
- Orthogonality, 28
- Overfitting, 79
  
- Parameter sharing, 176
- Partial derivative, 58
- Partition function, 101, 263, 352, 397
- PCA, *see* principal components analysis
- PCD, *see* stochastic maximum likelihood
- Perceptron, 13, 17
- Persistent contrastive divergence, *see* stochastic maximum likelihood
- Pooling, 173, 402
- Positive definite, 63
- Positive phase, 360
- Pre-training, 308
- Precision (of a normal distribution), 45, 47
- Predictive sparse decomposition, 285, 296
- Preprocessing, 242
- Principal components analysis, 32, 244, 372
- Principle components analysis, 91
- Probabilistic max pooling, 402
- Probability density function, 38
- Probability distribution, 37
- Probability mass function, 38
- Product rule of probability, *see* chain rule of probability
- PSD, *see* predictive sparse decomposition
- Pseudolikelihood, 365
  
- Random variable, 37
- Ratio matching, 368
- RBM, *see* restricted Boltzmann machine
- Receptive field, 177
- Recurrent network, 17
- Regression, 71
- Regularization, 131
- Representation learning, 5
- Restricted Boltzmann machine, 277, 372, 382, 383, 398, 400–402
- Ridge regression, 133
- Risk, 156
  
- Scalar, 20
- Score matching, 367



Second derivative, 62  
 Second derivative test, 63  
 Self-information, 42  
 Separable convolution, 190  
 Separation (probabilistic modeling), 266  
 SGD, *see* stochastic gradient descent, *see* stochastic gradient descent  
 Shannon entropy, 42, 379  
 Sigmoid, *see* logistic sigmoid  
 Sigmoid belief network, 17  
 Singular value decomposition, 30, 140  
 SML, *see* stochastic maximum likelihood  
 Softmax, 111  
 Softplus, 48  
 Spam detection, 5  
 Sparse coding, 292, 372  
 spectral radius, 211  
 Sphering, *see* Whitening, 244  
 Spike and slab restricted Boltzmann machine, 401  
 Square matrix, 26  
 ssRBM, *see* spike and slab restricted Boltzmann machine  
 Standard deviation, 41  
 Statistic, 83  
 Steepest descent, *see* gradient descent  
 Stochastic gradient descent, 158, 165, 398  
 Stochastic maximum likelihood, 364, 397, 398  
 Stochastic pooling, 154  
 Structure learning, 273  
 Structured output, 71  
 Structured probabilistic model, 255  
 Sum rule of probability, 39  
 Surrogate loss function, 157  
 SVD, *see* singular value decomposition  
 Symmetric matrix, 28  
 Tangent plane, 340  
 Tensor, 21  
 Tiled convolution, 186  
 Toeplitz matrix, 175  
 Trace operator, 31  
 Transcription, 71  
 Transfer learning, 315  
 Transpose, 22  
 Triangle inequality, 27  
 Triangulated graph, *see* chordal graph  
 Unbiased, 84  
 Underfitting, 79  
 Undirected model, 261  
 Uniform distribution, 38  
 Unit norm, 28  
 Unnormalized probability distribution, 262  
 Unsupervised pre-training, 308  
 V-structure, *see* explaining away  
 Vapnik-Chervonenkis dimension, 78  
 Variance, 41  
 Variational derivatives, *see* functional derivatives  
 Variational free energy, *see* evidence lower bound  
 VC dimension, *see* Vapnik-Chervonenkis dimension  
 Vector, 20  
 Visible layer, 9  
 Viterbi decoding, 226  
 Weight decay, 133  
 Whitening, 244  
 ZCA, *see* zero-phase components analysis  
 zero-data learning, 319  
 Zero-phase components analysis, 244  
 zero-shot learning, 319