

Deep Learning

Yoshua Bengio
Ian J. Goodfellow
Aaron Courville

January 1, 2015

Table of Contents

1	Deep Learning for AI	2
1.1	Who should read this book?	10
1.2	Machine Learning	11
1.3	Historical Perspective and Neural Networks	14
1.4	Recent Impact of Deep Learning Research	15
1.5	Challenges for Future Research	17
2	Linear algebra	20
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
3	Probability and Information Theory	35
3.1	Why probability?	35
3.2	Random variables	37
3.3	Probability distributions	37
3.3.1	Discrete variables and probability mass functions	37
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	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
4	Numerical Computation	56
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
5	Machine Learning Basics	70
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	81
5.5.1	Point Estimation	81
5.5.2	Bias	84
5.5.3	Variance	84
5.5.4	Trading off Bias and Variance and the Mean Squared Error	85
5.5.5	Consistency	86
5.6	Maximum likelihood estimation	86
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	90

5.10	Weakly supervised learning	92
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
6	Feedforward Deep Networks	102
6.1	Formalizing and Generalizing Neural Networks	102
6.2	Parametrizing a Learned Predictor	105
6.2.1	Family of Functions	105
6.2.2	Loss Function and Conditional Log-Likelihood	106
6.2.3	Training Criterion and Regularizer	111
6.2.4	Optimization Procedure	112
6.3	Flow Graphs and Back-Propagation	113
6.3.1	Chain Rule	114
6.3.2	Back-Propagation in a General Flow Graph	116
6.4	Universal Approximation Properties and Depth	120
6.5	Feature / Representation Learning	122
6.6	Piecewise Linear Hidden Units	124
6.7	Historical Notes	125
7	Regularization	126
7.1	Classical Regularization: Parameter Norm Penalty	127
7.1.1	L^2 Parameter Regularization	128
7.1.2	L^1 Regularization	130
7.1.3	L^∞ Regularization	132
7.2	Classical Regularization as Constrained Optimization	132
7.3	Regularization from a Bayesian Perspective	134
7.4	Early Stopping as a Form of Regularization	134
7.5	Regularization and Under-Constrained Problems	139
7.6	Parameter Sharing	140
7.7	Sparse Representations	140
7.8	Dataset Augmentation	140
7.9	Classical Regularization as Noise Robustness	141
7.10	Semi-Supervised Training	141
7.11	Unsupervised Pretraining	142
7.11.1	Pretraining Protocol	142
7.12	Bagging and Other Ensemble Methods	144
7.13	Dropout	146
7.14	Multi-Task Learning	149
8	Optimization for training deep models	150
8.1	Optimization for model training	150
8.1.1	Early Stopping	150
8.1.2	Plateaus, saddle points, and other flat regions	150

8.1.3	Cliffs and Exploding Gradients	150
8.1.4	Vanishing and Exploding Gradients - An Introduction to the Issue of Learning Long-Term Dependencies	153
8.2	Optimization algorithms	156
8.2.1	Approximate Natural Gradient and Second-Order Methods	156
8.2.2	Optimization strategies and meta-algorithms	156
8.2.3	Coordinate descent	156
8.2.4	Greedy supervised pre-training	157
8.3	Hints and Curriculum Learning	157
9	Structured Probabilistic Models: A Deep Learning Perspective	158
9.1	The Challenge of Unstructured Modeling	159
9.2	A Graphical Syntax for Describing Model Structure	161
9.2.1	Directed Models	162
9.2.2	Undirected Models	163
9.2.3	The Partition Function	164
9.2.4	Energy-Based Models	166
9.2.5	Separation and D-Separation	167
9.2.6	Operations on a Graph	169
9.2.7	Factor Graphs	169
9.3	Advantages of Structured Modeling	171
9.4	Learning about Dependencies	171
9.4.1	Latent Variables Versus Structure Learning	171
9.4.2	Latent Variables for Feature Learning	172
9.5	Markov Chain Monte Carlo Methods	173
9.6	Inference and Approximate Inference Over Latent Variables	174
9.7	The Deep Learning Approach to Structured Probabilistic Modeling	176
9.7.1	Example: The Restricted Boltzmann Machine	177
10	Unsupervised and Transfer Learning	179
10.1	Auto-Encoders	180
10.1.1	Regularized Auto-Encoders	181
10.1.2	Representational Power, Layer Size and Depth	184
10.1.3	Reconstruction Distribution	185
10.2	Linear Factor Models	186
10.2.1	Probabilistic PCA and Factor Analysis	186
10.2.2	Manifold Interpretation of PCA and Linear Auto-Encoders	188
10.2.3	ICA	190
10.2.4	Sparse Coding as a Generative Model	191
10.3	RBMs	192
10.4	Greedy Layerwise Unsupervised Pre-Training	192
10.5	Transfer Learning and Domain Adaptation	193

11 Convolutional Networks	199
11.1 The convolution operation	199
11.2 Motivation	201
11.3 Pooling	204
11.4 Variants of the basic convolution function	209
11.5 Data types	214
11.6 Efficient convolution algorithms	216
11.7 Deep learning history	216
12 Sequence Modeling: Recurrent and Recursive Nets	217
12.1 Unfolding Flow Graphs and Sharing Parameters	217
12.2 Recurrent Neural Networks	219
12.2.1 Computing the gradient in a recurrent neural network	221
12.2.2 Recurrent Networks as Generative Directed Acyclic Models	223
12.2.3 RNNs to represent conditional probability distributions	225
12.3 Bidirectional RNNs	227
12.4 Recursive Neural Networks	229
12.5 Auto-Regressive Networks	230
12.5.1 Logistic Auto-Regressive Networks	231
12.5.2 Neural Auto-Regressive Networks	232
12.5.3 NADE	234
12.6 Facing the Challenge of Long-Term Dependencies	235
12.6.1 Echo State Networks: Choosing Weights to Make Dynamics Barely Contractive	235
12.6.2 Combining Short and Long Paths in the Unfolded Flow Graph	237
12.6.3 Leaky Units and a Hierarchy Different Time Scales	238
12.6.4 The Long-Short-Term-Memory Architecture and Other Gated RNNs	239
12.6.5 Deep RNNs	241
12.6.6 Better Optimization	243
12.6.7 Clipping Gradients	244
12.6.8 Regularizing to Encourage Information Flow	245
12.6.9 Organizing the State at Multiple Time Scales	245
12.7 Handling temporal dependencies with n-grams, HMMs, CRFs and other graphical models	246
12.7.1 N-grams	246
12.7.2 Efficient Marginalization and Inference for Temporally Structured Outputs by Dynamic Programming	247
12.7.3 HMMs	252
12.7.4 CRFs	254
12.8 Combining Neural Networks and Search	256
12.8.1 Approximate Search	257

13 The Manifold Perspective on Auto-Encoders	261
13.1 Manifold Learning via Regularized Auto-Encoders	269
13.2 Probabilistic Interpretation of Reconstruction Error as Log-Likelihood	272
13.3 Sparse Representations	273
13.3.1 Sparse Auto-Encoders	274
13.3.2 Predictive Sparse Decomposition	276
13.4 Denoising Auto-Encoders	277
13.4.1 Learning a Vector Field that Estimates a Gradient Field	279
13.4.2 Turning the Gradient Field into a Generative Model	281
13.5 Contractive Auto-Encoders	284
13.6 Tangent Distance, Tangent-Prop, and Manifold Tangent Classifier	285
14 Distributed Representations: Disentangling the Underlying Factors	288
14.1 Causality and Semi-Supervised Learning	288
14.2 Assumption of Underlying Factors and Distributed Representation	290
14.3 Exponential Gain in Representational Efficiency from Distributed Representations	294
14.4 Exponential Gain in Representational Efficiency from Depth	295
14.5 Priors Regarding The Underlying Factors	298
15 Confronting the Partition Function	301
15.1 Estimating the partition function	301
15.1.1 Annealed importance sampling	303
15.1.2 Bridge sampling	306
15.1.3 Extensions	306
15.2 Stochastic maximum likelihood and contrastive divergence	307
15.3 Pseudolikelihood	314
15.4 Score matching and ratio matching	316
15.5 Denoising score matching	318
15.6 Noise-contrastive estimation	318
16 Approximate inference	321
16.1 Inference as optimization	321
16.2 Expectation maximization	323
16.3 MAP inference: Sparse coding as a probabilistic model	324
16.4 Variational inference and learning	325
16.4.1 Discrete latent variables	327
16.4.2 Calculus of variations	327
16.4.3 Continuous latent variables	329
16.5 Stochastic inference	329
16.6 Learned approximate inference	329

17 Deep generative models	330
17.1 Restricted Boltzmann machines	330
17.2 Deep belief networks	332
17.3 Deep Boltzmann machines	333
17.3.1 Interesting properties	333
17.3.2 Variational learning with SML	334
17.3.3 Layerwise pretraining	335
17.3.4 Multi-prediction deep Boltzmann machines	337
17.3.5 Centered deep Boltzmann machines	337
17.4 Boltzmann machines for real-valued data	337
17.4.1 Gaussian-Bernoulli RBMs	337
17.4.2 mcRBMs	338
17.4.3 Spike and slab restricted Boltzmann machines	338
17.5 Convolutional Boltzmann machines	338
17.6 Other Boltzmann machines	339
17.7 Directed generative nets	339
17.7.1 Variational autoencoders	339
17.7.2 Variational interpretation of PSD	339
17.7.3 Generative adversarial networks	339
17.8 A generative view of autoencoders	340
17.9 Generative stochastic networks	340
17.10 Methodological notes	340
18 Large scale deep learning	343
18.1 Fast CPU implementations	343
18.2 GPU implementations	343
18.3 Asynchronous parallel implementations	343
18.4 Dynamically structured nets	343
18.5 Model compression	344
19 Practical methodology	345
19.1 When to gather more data, control capacity, or change algorithms	345
19.2 Machine Learning Methodology 101	345
19.3 Manual hyperparameter tuning	345
19.4 Hyper-parameter optimization algorithms	345
19.5 Tricks of the Trade for Deep Learning	347
19.5.1 Debugging Back-Prop	347
19.5.2 Automatic Differentiation and Symbolic Manipulations of Flow Graphs	347
19.5.3 Momentum and Other Averaging Techniques as Cheap Second Order Methods	347

20 Applications	348
20.1 Computer vision	348
20.1.1 Preprocessing	349
20.1.2 Convolutional nets	354
20.2 Speech Recognition	354
20.3 Natural language processing and neural language models	354
20.3.1 Neural language models	354
20.4 Structured outputs	354
20.5 Other applications	354
Bibliography	355
Index	376

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: Pawel Chilinski.

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

Math background chapters: Ilya Sutskever, Vincent Vanhoucke, Johannes Roith,

Linear algebra: Guillaume Alain, Dustin Webb, David Warde-Farley, Pierre Luc Carrier, Li Yao, Thomas Rohée, Colby Toland, Amjad Almahairi, Sergey Oreshkov,

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

Numerical: Meire Fortunato, Jurgen Van Gael. Dustin Webb

ML: Dzmitry Bahdanau Kelvin Xu

MLPs: Jurgen Van Gael

Convolutional nets: Guillaume Alain, David Warde-Farley, Mehdi Mirza, Caglar Gulcehre.

Unsupervised: Kelvin Xu

Partition function: Sam Bowman.

Graphical models: Kelvin Xu

RNNs: Kelvin Xu Dmitriy Serdyuk

We also want to thank Jason Yosinski and Nicolas Chapados for contributing figures (as noted in the captions).

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

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. 279
- Alain, G. and Bengio, Y. (2013). What regularized auto-encoders learn from the data generating distribution. In *ICLR'2013*. also arXiv report 1211.4246. 279, 281
- Amari, S. (1997). Neural learning in structured parameter spaces - natural Riemannian gradient. In *Advances in Neural Information Processing Systems*, pages 127–133. MIT Press. 113
- Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. Technical report, arXiv preprint arXiv:1409.0473. 10
- 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, 254
- Baldi, P. and Brunak, S. (1998). *Bioinformatics, the Machine Learning Approach*. MIT Press. 256
- Baldi, P. and Sadowski, P. J. (2013). Understanding dropout. In *Advances in Neural Information Processing Systems 26*, pages 2814–2822. 149
- 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. 228
- Barron, A. E. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. on Information Theory*, **39**, 930–945. 121
- Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. Oxford University Press. 187
- Basilevsky, A. (1994). *Statistical Factor Analysis and Related Methods: Theory and Applications*. Wiley. 187
- 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. 252
- Becker, S. and Hinton, G. (1992). A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, **355**, 161–163. 300

- Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, **15**(6), 1373–1396. 98, 265
- 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. 232
- Bengio, Y. (1991). *Artificial Neural Networks and their Application to Sequence Recognition*. Ph.D. thesis, McGill University, (Computer Science), Montreal, Canada. 237, 256
- Bengio, Y. (1993). A connectionist approach to speech recognition. *International Journal on Pattern Recognition and Artificial Intelligence*, **7**(4), 647–668. 254
- Bengio, Y. (1999a). Markovian models for sequential data. *Neural Computing Surveys*, **2**, 129–162. 254
- Bengio, Y. (1999b). Markovian models for sequential data. *Neural Computing Surveys*, **2**, 129–162. 256
- Bengio, Y. (2009). *Learning deep architectures for AI*. Now Publishers. 18, 95, 122
- Bengio, Y. (2011). Deep learning of representations for unsupervised and transfer learning. In *JMLR W&CP: Proc. Unsupervised and Transfer Learning*. 18
- Bengio, Y. and Bengio, S. (2000b). Modeling high-dimensional discrete data with multi-layer neural networks. In *NIPS'99*, pages 400–406. MIT Press. 232, 234, 235
- Bengio, Y. and Delalleau, O. (2009). Justifying and generalizing contrastive divergence. *Neural Computation*, **21**(6), 1601–1621. 310
- Bengio, Y. and Frasconi, P. (1996). Input/Output HMMs for sequence processing. *IEEE Transactions on Neural Networks*, **7**(5), 1231–1249. 256
- 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. 123
- Bengio, Y. and Monperrus, M. (2005). Non-local manifold tangent learning. In *NIPS'04*, pages 129–136. MIT Press. 97, 266
- 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. 254, 256
- 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). 155, 243
- Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Tr. Neural Nets*. 155, 156, 235, 241, 243

- 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. 256
- Bengio, Y., Ducharme, R., and Vincent, P. (2001a). A neural probabilistic language model. In *NIPS'00*, pages 932–938. MIT Press. 14
- Bengio, Y., Ducharme, R., and Vincent, P. (2001b). A neural probabilistic language model. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *NIPS'2000*, pages 932–938. MIT Press. 267, 269
- Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, **3**, 1137–1155. 267, 269
- 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, 265
- Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007). Greedy layer-wise training of deep networks. In *NIPS'2006*. 15, 142, 192
- Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *ICML'09*. 113
- Bengio, Y., Yao, L., Alain, G., and Vincent, P. (2013a). Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems 26 (NIPS'13)*. 282
- Bengio, Y., Courville, A., and Vincent, P. (2013b). Representation learning: A review and new perspectives. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, **35**(8), 1798–1828. 298, 299, 339
- Bengio, Y., Thibodeau-Laufer, E., Alain, G., and Yosinski, J. (2014). Deep generative stochastic networks trainable by backprop. In *Proceedings of the 30th International Conference on Machine Learning (ICML'14)*. 282, 283
- Bennett, C. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, **22**(2), 245–268. 306
- Berglund, M. and Raiko, T. (2013). Stochastic gradient estimate variance in contrastive divergence and persistent contrastive divergence. *CoRR*, **abs/1312.6002**. 313
- Bergstra, J. (2011). *Incorporating Complex Cells into Neural Networks for Pattern Classification*. Ph.D. thesis, Université de Montréal. 183
- 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. 315
- Bishop, C. M. (1994). Mixture density networks. 109

- 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*. 230
- 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. 14, 95, 106
- Bottou, L. (1991). *Une approche théorique de l'apprentissage connexioniste; applications à la reconnaissance de la parole*. Ph.D. thesis, Université de Paris XI. 256
- Bottou, L. (2011). From machine learning to machine reasoning. Technical report, arXiv.1102.1808. 229, 230
- 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. 256
- 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. 247, 254, 255, 256, 257, 258, 260
- Bourlard, H. and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, **59**, 291–294. 180
- 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. 256
- 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. 256
- Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA. 65
- Brand, M. (2003). Charting a manifold. In *NIPS'2002*, pages 961–968. MIT Press. 98, 265
- Breiman, L. (1994). Bagging predictors. *Machine Learning*, **24**(2), 123–140. 144
- 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. 254
- Carreira-Perpiñán, 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. 310

- 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. 13, 98, 261
- 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. 246, 247
- Cho, K., van Merriënboer, B., Gulcehre, C., 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)*. 241
- Ciresan, D., Meier, U., Masci, J., and Schmidhuber, J. (2012). Multi-column deep neural network for traffic sign classification. *Neural Networks*, 32, 333–338. 15, 122
- 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)*. 350
- Collobert, R. (2004). *Large Scale Machine Learning*. Ph.D. thesis, Université de Paris VI, LIP6. 106
- Comon, P. (1994). Independent component analysis - a new concept? *Signal Processing*, 36, 287–314. 190
- Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, 20, 273–297. 14, 95
- Coupric, C., Farabet, C., Najman, L., and LeCun, Y. (2013). Indoor semantic segmentation using depth information. In *International Conference on Learning Representations (ICLR2013)*. 15, 122
- Courville, A., Bergstra, J., and Bengio, Y. (2011). Unsupervised models of images by spike-and-slab RBMs. In *ICML'11*. 160
- Cover, T. (2006). *Elements of Information Theory*. Wiley-Interscience. 42
- Crick, F. H. C. and Mitchison, G. (1983). The function of dream sleep. *Nature*, 304, 111–114. 309
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2, 303–314. 297
- Dahl, G. E., Ranzato, M., Mohamed, A., and Hinton, G. E. (2010). Phone recognition with the mean-covariance restricted Boltzmann machine. In *NIPS'2010*. 15
- Dauphin, Y. and Bengio, Y. (2013a). Big neural networks waste capacity. In *ICLR'2013 workshops track (oral presentation)*, *arXiv: 1301.3583*. 17
- Dauphin, Y. and Bengio, Y. (2013b). Stochastic ratio matching of RBMs for sparse high-dimensional inputs. In *NIPS26*. NIPS Foundation. 318

- 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
- 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. 348
- Delalleau, O. and Bengio, Y. (2011). Shallow vs. deep sum-product networks. In *NIPS*. 122, 296, 297
- 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. 248
- Deng, L., Seltzer, M., Yu, D., Acero, A., Mohamed, A., and Hinton, G. (2010). Binary coding of speech spectrograms using a deep auto-encoder. In *Interspeech 2010*, Makuhari, Chiba, Japan. 15
- 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. 338
- Desjardins, G., Courville, A., and Bengio, Y. (2011). On tracking the partition function. In *NIPS'2011*. 307
- Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., and Makhoul, J. (2014). Fast and robust neural network joint models for statistical machine translation. In *Proc. ACL'2014*. 10
- Do, T.-M.-T. and Artières, T. (2010). Neural conditional random fields. In *International Conference on Artificial Intelligence and Statistics*, pages 177–184. 248
- 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, 265
- Doya, K. (1993). Bifurcations of recurrent neural networks in gradient descent learning. *IEEE Transactions on Neural Networks*, **1**, 75–80. 156, 235
- 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. 106
- 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)*. 9, 122
- El Hihi, S. and Bengio, Y. (1996). Hierarchical recurrent neural networks for long-term dependencies. In *NIPS 8*. MIT Press. 242, 245, 246
- ElHihi, S. and Bengio, Y. (1996). Hierarchical recurrent neural networks for long-term dependencies. In *NIPS'1995*. 238

- Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *JMLR*, **11**, 625–660. [18](#)
- 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. [276](#)
- Farabet, C., Couprie, C., Najman, L., and LeCun, Y. (2013a). Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. [15](#), [122](#)
- 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. [248](#)
- 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. [196](#)
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188. [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*. [229](#), [230](#)
- 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. [230](#)
- Frey, B. J. (1998). *Graphical models for machine learning and digital communication*. MIT Press. [231](#)
- 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](#)
- 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. [121](#)
- Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *AISTATS’2010*. [15](#)
- Glorot, X., Bordes, A., and Bengio, Y. (2011a). Deep sparse rectifier neural networks. In *AISTATS’2011*. [15](#), [106](#), [275](#)
- 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)*. [124](#), [275](#)
- Glorot, X., Bordes, A., and Bengio, Y. (2011c). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML’2011*. [193](#), [275](#)
- Gong, S., McKenna, S., and Psarrou, A. (2000). *Dynamic Vision: From Images to Face Recognition*. Imperial College Press. [264](#), [267](#)

- Goodfellow, I., Le, Q., Saxe, A., and Ng, A. (2009). Measuring invariances in deep networks. In *NIPS'2009*, pages 646–654. [183](#), [275](#)
- 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*. [192](#)
- Goodfellow, I. J. (2010). Technical report: Multidimensional, downsampled convolution for autoencoders. Technical report, Université de Montréal. [213](#)
- 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*. [9](#), [18](#), [122](#), [194](#)
- Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y. (2013a). Maxout networks. In *ICML'2013*. [15](#)
- Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y. (2013b). Maxout networks. In S. Dasgupta and D. McAllester, editors, *ICML'13*, pages 1319–1327. [124](#), [148](#), [350](#)
- Goodfellow, I. J., Mirza, M., Courville, A., and Bengio, Y. (2013c). Multi-prediction deep Boltzmann machines. In *NIPS26*. NIPS Foundation. [316](#), [335](#), [336](#)
- Gouws, S., Bengio, Y., and Corrado, G. (2014). Bilbowa: Fast bilingual distributed representations without word alignments. Technical report, arXiv:1410.2455. [196](#)
- Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks*. Studies in Computational Intelligence. Springer. [227](#), [240](#), [241](#), [247](#)
- Graves, A. (2013). Generating sequences with recurrent neural networks. Technical report, arXiv preprint arXiv:1308.0850. [110](#), [240](#), [242](#)
- Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, **18**(5), 602–610. [227](#)
- 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. [227](#)
- 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. [247](#)
- 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. [227](#)
- Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *ICASSP'2013*, pages 6645–6649. IEEE. [228](#), [240](#), [241](#)

- Gulcehre, C. and Bengio, Y. (2013). Knowledge matters: Importance of prior information for optimization. In *International Conference on Learning Representations (ICLR'2013)*. 18
- 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)*. 318
- 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. 256
- 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. 122, 297
- Håstad, J. and Goldmann, M. (1991). On the power of small-depth threshold circuits. *Computational Complexity*, 1, 113–129. 122, 297
- Henaff, M., Jarrett, K., Kavukcuoglu, K., and LeCun, Y. (2011). Unsupervised learning of sparse features for scalable audio classification. In *ISMIR'11*. 276
- Herauld, 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. 190
- Hermann, K. M. and Blunsom, P. (2014). Multilingual Distributed Representations without Word Alignment. In *Proceedings of ICLR*. 10
- Hinton, G., Deng, L., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. (2012a). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6), 82–97. 10, 15
- Hinton, G. E. (2000). Training products of experts by minimizing contrastive divergence. Technical Report GCNU TR 2000-004, Gatsby Unit, University College London. 309
- Hinton, G. E. and Roweis, S. (2003). Stochastic neighbor embedding. In *NIPS'2002*. 265
- Hinton, G. E. and Salakhutdinov, R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, 313, 504–507. 142
- Hinton, G. E. and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786), 504–507. 185, 192, 193
- Hinton, G. E. and Zemel, R. S. (1994). Autoencoders, minimum description length, and Helmholtz free energy. In *NIPS'1993*. 180
- Hinton, G. E., Osindero, S., and Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18, 1527–1554. 15, 142, 192, 193, 332
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2012b). Improving neural networks by preventing co-adaptation of feature detectors. Technical report, arXiv:1207.0580. 133

- Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, T.U. Munich. [155](#), [235](#), [243](#)
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, **9**(8), 1735–1780. [240](#), [241](#)
- 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. [241](#)
- Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**, 359–366. [297](#)
- Hsu, F.-H. (2002). *Behind Deep Blue: Building the Computer That Defeated the World Chess Champion*. Princeton University Press, Princeton, NJ, USA. [2](#)
- 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. [315](#)
- Hyotyniemi, H. (1996). Turing machines are recurrent neural networks. In *STeP’96*, pages 13–24. [219](#)
- Hyvärinen, A. (1999). Survey on independent component analysis. *Neural Computing Surveys*, **2**, 94–128. [190](#)
- Hyvärinen, A. (2005). Estimation of non-normalized statistical models using score matching. *Journal of Machine Learning Research*, **6**, 695–709. [316](#)
- Hyvärinen, A. (2007a). Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Transactions on Neural Networks*, **18**, 1529–1531. [317](#)
- Hyvärinen, A. (2007b). Some extensions of score matching. *Computational Statistics and Data Analysis*, **51**, 2499–2512. [317](#)
- Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. Wiley-Interscience. [190](#)
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixture of local experts. *Neural Computation*, **3**, 79–87. [109](#)
- Jaeger, H. (2003). Adaptive nonlinear system identification with echo state networks. In *Advances in Neural Information Processing Systems 15*. [236](#)
- Jaeger, H. (2007a). Discovering multiscale dynamical features with hierarchical echo state networks. Technical report, Jacobs University. [242](#)
- Jaeger, H. (2007b). Echo state network. *Scholarpedia*, **2**(9), 2330. [235](#)
- Jaeger, H. and Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, **304**(5667), 78–80. [235](#)
- 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. [289](#)

- Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009a). What is the best multi-stage architecture for object recognition? In *ICCV'09*. 106, 276
- Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009b). What is the best multi-stage architecture for object recognition? In *Proc. International Conference on Computer Vision (ICCV'09)*, pages 2146–2153. IEEE. 124
- Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, **78**, 2690–2693. 306
- 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. 246
- Jordan, M. I. (1998). *Learning in Graphical Models*. Kluwer, Dordrecht, Netherlands. 14
- Juang, B. H. and Katagiri, S. (1992). Discriminative learning for minimum error classification. *IEEE Transactions on Signal Processing*, **40**(12), 3043–3054. 254
- Jutten, C. and Herault, J. (1991). Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture. *Signal Processing*, **24**, 1–10. 190
- 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. 246
- 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. 183
- 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. 276
- Kavukcuoglu, K., Ranzato, M.-A., Fergus, R., and LeCun, Y. (2009). Learning invariant features through topographic filter maps. In *CVPR'2009*. 276
- 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*. 276
- Kindermann, R. (1980). *Markov Random Fields and Their Applications (Contemporary Mathematics ; V. 1)*. American Mathematical Society. 164
- Kingma, D. and LeCun, Y. (2010). 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. 318
- Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 267, 268
- Klementiev, A., Titov, I., and Bhattacharai, B. (2012). Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*. 196

- Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press. 172, 173, 252
- Koren, Y. (2009). 1 the bellkor solution to the netflix grand prize. 146
- Koutnik, J., Greff, K., Gomez, F., and Schmidhuber, J. (2014). A clockwork RNN. In *ICML'2014*. 242, 246
- 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*. 297
- Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto. 160
- Krizhevsky, A., Sutskever, I., and Hinton, G. (2012a). ImageNet classification with deep convolutional neural networks. In *NIPS'2012*. 9, 15, 122, 275
- Krizhevsky, A., Sutskever, I., and Hinton, G. (2012b). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (NIPS'2012)*. 71
- 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. 248, 254
- Lake, B., Salakhutdinov, R., and Tenenbaum, J. (2013). One-shot learning by inverting a compositional causal process. In *NIPS'2013*. 18
- 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. 217, 237
- Larochelle, H. and Bengio, Y. (2008). Classification using discriminative restricted Boltzmann machines. In *ICML'2008*. 183
- Larochelle, H. and Murray, I. (2011). The Neural Autoregressive Distribution Estimator. In *AISTATS'2011*. 230, 234
- Larochelle, H., Erhan, D., and Bengio, Y. (2008). Zero-data learning of new tasks. In *AAAI Conference on Artificial Intelligence*. 18, 196
- Le Roux, N. and Bengio, Y. (2010). Deep belief networks are compact universal approximators. *Neural Computation*, 22(8), 2192–2207. 297
- Le Roux, N., Manzagol, P.-A., and Bengio, Y. (2008). Topmoumoute online natural gradient algorithm. In *NIPS'07*. 113
- LeCun, Y. (1987). *Modèles connexionistes de l'apprentissage*. Ph.D. thesis, Université de Paris VI. 14, 180
- 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. 15

- 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. [14](#), [247](#), [254](#), [255](#), [256](#)
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998b). Gradient based learning applied to document recognition. *Proc. IEEE*. [15](#)
- Lee, H., Ekanadham, C., and Ng, A. (2008). Sparse deep belief net model for visual area V2. In *NIPS'07*. [183](#)
- 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. [338](#), [339](#)
- Leprieur, H. and Haffner, P. (1995). Discriminant learning with minimum memory loss for improved non-vocabulary rejection. In *EUROSPEECH'95*, Madrid, Spain. [254](#)
- 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. [237](#)
- Linde, N. (1992). The machine that changed the world, episode 3. Documentary miniseries. [3](#)
- 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)*. [330](#)
- Lovelace, A. (1842). Notes upon L. F. Menabrea's "Sketch of the Analytical Engine invented by Charles Babbage". [2](#)
- Lowerre, B. (1976). *The Harpy Speech Recognition System*. Ph.D. thesis. [248](#), [253](#), [258](#)
- Lukoševičius, M. and Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, **3**(3), 127–149. [235](#)
- 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*. [317](#)
- 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. [235](#)
- 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. [313](#), [317](#)
- Martens, J. and Medabalimi, V. (2014). On the expressive efficiency of sum product networks. *arXiv preprint arXiv:1411.7717*. [297](#)
- Martens, J. and Sutskever, I. (2011). Learning recurrent neural networks with Hessian-free optimization. In *Proc. ICML'2011*. ACM. [243](#)
- 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. [315](#)

- 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. 256
- McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London. 107
- 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. 9, 18, 122, 194
- Mesnil, G., Rifai, S., Dauphin, Y., Bengio, Y., and Vincent, P. (2012). Surfing on the manifold. Learning Workshop, Snowbird. 281
- Mikolov, T. (2012). *Statistical Language Models based on Neural Networks*. Ph.D. thesis, Brno University of Technology. 110, 244
- Mikolov, T., Le, Q. V., and Sutskever, I. (2013). Exploiting similarities among languages for machine translation. Technical report, arXiv:1309.4168. 196
- Minka, T. (2005). Divergence measures and message passing. *Microsoft Research Cambridge UK Tech Rep MSRTR2005173*, 72(TR-2005-173). 303
- Minsky, M. L. and Papert, S. A. (1969). *Perceptrons*. MIT Press, Cambridge. 14
- 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. 320
- Montúfar, G. (2014). Universal approximation depth and errors of narrow belief networks with discrete units. *Neural Computation*, 26. 297
- 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. 297
- Montufar, G. and Morton, J. (2014). When does a mixture of products contain a product of mixtures? *SIAM Journal on Discrete Mathematics (SIDMA)*. 295
- Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. (2014). On the number of linear regions of deep neural networks. In *NIPS'2014*. 294, 297, 298
- 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. 3
- Moser, M. C. (1992). The induction of multiscale temporal structure. In *NIPS'91*, pages 275–282, San Mateo, CA. Morgan Kaufmann. 238, 246
- Murphy, K. P. (2012). *Machine Learning: a Probabilistic Perspective*. MIT Press, Cambridge, MA, USA. 108

- Murray, B. U. I. and Larochelle, H. (2014). A deep and tractable density estimator. In *ICML'2014*. 110, 234, 235
- 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. 254
- Nair, V. and Hinton, G. (2010a). Rectified linear units improve restricted Boltzmann machines. In *ICML'2010*. 106, 275
- Nair, V. and Hinton, G. E. (2010b). Rectified linear units improve restricted Boltzmann machines. In L. Bottou and M. Littman, editors, *Proceedings of the Twenty-seventh International Conference on Machine Learning (ICML-10)*, pages 807–814. ACM. 15
- Narayanan, H. and Mitter, S. (2010). Sample complexity of testing the manifold hypothesis. In *NIPS'2010*. 13, 98, 261
- Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics. Springer. 149
- Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, **11**(2), 125–139. 305, 306
- Neal, R. M. (2005). Estimating ratios of normalizing constants using linked importance sampling. 306, 307
- Niranjan, M. and Fallside, F. (1990). Neural networks and radial basis functions in classifying static speech patterns. *Computer Speech and Language*, **4**, 275–289. 106
- 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. 182, 183, 300
- 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. 274
- Park, H., Amari, S.-I., and Fukumizu, K. (2000). Adaptive natural gradient learning algorithms for various stochastic models. *Neural Networks*, **13**(7), 755 – 764. 113
- Pascanu, R. (2014). *On recurrent and deep networks*. Ph.D. thesis, Université de Montréal. 152, 153
- Pascanu, R. and Bengio, Y. (2012). On the difficulty of training recurrent neural networks. Technical Report arXiv:1211.5063, Université de Montréal. 110
- Pascanu, R. and Bengio, Y. (2013). Revisiting natural gradient for deep networks. Technical report, arXiv:1301.3584. 113
- Pascanu, R., Mikolov, T., and Bengio, Y. (2013a). On the difficulty of training recurrent neural networks. In *ICML'2013*. 110, 156, 235, 238, 244, 245, 246
- 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. 122

- Pascanu, R., Gulcehre, C., Cho, K., and Bengio, Y. (2014a). How to construct deep recurrent neural networks. In *ICLR'2014*. 148
- Pascanu, R., Gulcehre, C., Cho, K., and Bengio, Y. (2014b). How to construct deep recurrent neural networks. In *ICLR'2014*. 240, 242, 297
- 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*. 294
- 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. 162
- 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. 339
- Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, 46(1), 77–105. 229
- Poon, H. and Domingos, P. (2011). Sum-product networks: A new deep architecture. In *UAI'2011*, Barcelona, Spain. 122, 296, 297
- Powell, M. (1987). Radial basis functions for multivariable interpolation: A review. 106
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286. 252
- Rabiner, L. R. and Juang, B. H. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, pages 257–285. 217, 252
- Raiko, T., Yao, L., Cho, K., and Bengio, Y. (2014). Iterative neural autoregressive distribution estimator (NADE-k). Technical report, arXiv preprint arXiv:1406.1485. 234
- 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*. 15, 142, 192, 275
- Ranzato, M., Boureau, Y., and LeCun, Y. (2008). Sparse feature learning for deep belief networks. In *NIPS'2007*. 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)*. 18, 196
- Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. (2011a). Contractive auto-encoders: Explicit invariance during feature extraction. In *ICML'2011*. 284

- 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)*. 183
- Rifai, S., Mesnil, G., Vincent, P., Muller, X., Bengio, Y., Dauphin, Y., and Glorot, X. (2011c). Higher order contractive auto-encoder. In *ECML PKDD*. 284
- Rifai, S., Dauphin, Y., Vincent, P., Bengio, Y., and Muller, X. (2011d). The manifold tangent classifier. In *NIPS'2011*. 286, 287
- Rifai, S., Bengio, Y., Dauphin, Y., and Vincent, P. (2012). A generative process for sampling contractive auto-encoders. In *ICML'2012*. 281
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**, 386–408. 14
- Rosenblatt, F. (1962). *Principles of Neurodynamics*. Spartan, New York. 14
- Roweis, S. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**(5500). 98, 265
- Rumelhart, D., Hinton, G., and Williams, R. (1986a). Learning representations by back-propagating errors. *Nature*, **323**, 533–536. 14
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986b). Learning representations by back-propagating errors. *Nature*, **323**, 533–536. 102, 217
- Rumelhart, D. E., McClelland, J. L., and the PDP Research Group (1986c). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1. MIT Press, Cambridge. 102
- 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. 193, 333, 335
- 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. 337
- 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. 306
- Schmidhuber, J. (1992). Learning complex, extended sequences using the principle of history compression. *Neural Computation*, **4**(2), 234–242. 15, 242
- Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, **10**, 1299–1319. 98, 265
- Schölkopf, B., Burges, C. J. C., and Smola, A. J. (1999). *Advances in Kernel Methods — Support Vector Learning*. MIT Press, Cambridge, MA. 14, 106, 122
- Schuster, M. and Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, **45**(11), 2673–2681. 227

- Schölkopf, B. and Smola, A. (2002). *Learning with kernels*. MIT Press. 95
- Seide, F., Li, G., and Yu, D. (2011). Conversational speech transcription using context-dependent deep neural networks. In *Interspeech 2011*, pages 437–440. 15
- 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. 15, 122
- 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
- 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. 219
- Siegelmann, H. and Sontag, E. (1991). Turing computability with neural nets. *Applied Mathematics Letters*, **4**(6), 77–80. 219
- 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. 156
- 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*. 286, 287
- Simard, P. Y., LeCun, Y., and Denker, J. (1993). Efficient pattern recognition using a new transformation distance. In *NIPS’92*. 285
- 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**. 285
- 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. 167, 177
- 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*. 230
- 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)*. 230
- 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*. 230
- 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*. 230

- Solla, S. A., Levin, E., and Fleisher, M. (1988). Accelerated learning in layered neural networks. *Complex Systems*, **2**, 625–639. 108
- Srivastava, N. and Salakhutdinov, R. (2012). Multimodal learning with deep Boltzmann machines. In *NIPS'2012*. 197
- 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. 146, 148, 149, 335
- 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. 248
- Sutskever, I. (2012). *Training Recurrent Neural Networks*. Ph.D. thesis, Departement of computer science, University of Toronto. 236, 243
- 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. 312
- Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *ICML*. 236, 243
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. Technical report, arXiv preprint arXiv:1409.3215. 10, 240, 241
- 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. 318
- 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 preprint arXiv:1409.4842. 9
- Tenenbaum, J., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**(5500), 2319–2323. 98, 265
- Tibshirani, R. J. (1995). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **58**, 267–288. 132
- 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. 313
- Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal components analysis. *Journal of the Royal Statistical Society B*, **61**(3), 611–622. 187, 188
- Uribe, B., Murray, I., and Larochelle, H. (2013). Rnade: The real-valued neural autoregressive density-estimator. In *NIPS'2013*. 233, 234
- Utgoff, P. E. and Stracuzzi, D. J. (2002). Many-layered learning. *Neural Computation*, **14**, 2497–2539. 15

- van der Maaten, L. and Hinton, G. E. (2008). Visualizing data using t-SNE. *J. Machine Learning Res.*, **9**, 265, 268
- 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. (2011). A connection between score matching and denoising autoencoders. *Neural Computation*, **23**(7), 1661–1674. 282, 318
- Vincent, P. and Bengio, Y. (2003). Manifold Parzen windows. In *NIPS'2002*. MIT Press. 265
- Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *ICML 2008*. 277
- 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**. 277
- Wager, S., Wang, S., and Liang, P. (2013). Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems 26*, pages 351–359. 149
- 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. 217
- Wan, L., Zeiler, M., Zhang, S., LeCun, Y., and Fergus, R. (2013). Regularization of neural networks using dropconnect. In *ICML'2013*. 149
- Wang, S. and Manning, C. (2013). Fast dropout training. In *ICML'2013*. 149
- Warde-Farley, D., Goodfellow, I. J., Courville, A., and Bengio, Y. (2014). An empirical analysis of dropout in piecewise linear networks. In *ICLR'2014*. 149
- Weinberger, K. Q. and Saul, L. K. (2004). Unsupervised learning of image manifolds by semidefinite programming. In *CVPR'2004*, pages 988–995. 98, 265
- Weston, J., Ratle, F., and Collobert, R. (2008). Deep learning via semi-supervised embedding. In *ICML 2008*. 18
- 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. 230
- White, H. (1990). Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, **3**(5), 535–549. 121
- Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. In *1960 IRE WESCON Convention Record*, volume 4, pages 96–104. IRE, New York. 14

- Williams, C. K. I. and Rasmussen, C. E. (1996). Gaussian processes for regression. In *NIPS'95*, pages 514–520. MIT Press, Cambridge, MA. 122
- Wolpert, D. H. (1996). The lack of a priori distinction between learning algorithms. *Neural Computation*, **8**(7), 1341–1390. 121
- 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. 149
- Xu, L. and Jordan, M. I. (1996). On convergence properties of the EM algorithm for gaussian mixtures. *Neural Computation*, **8**, 129–151. 253
- Younes, L. (1998). On the convergence of Markovian stochastic algorithms with rapidly decreasing ergodicity rates. In *Stochastics and Stochastics Models*, pages 177–228. 313
- 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. 295
- Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *ECCV'14*. 6, 71
- 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. 112

Index

- L^p norm, 26
- Active constraint, 68
- AIS, *see* annealed importance sampling
- Almost everywhere, 52
- Ancestral sampling, 171
- Annealed importance sampling, 278, 309
- Approximate inference, 174
- Artificial intelligence, 2
- Asymptotically unbiased, 84
- Bagging, 144
- Bayes' rule, 51
- Bayesian network, *see* directed graphical model
- Bayesian probability, 37
- Belief network, *see* directed graphical model
- Bernoulli distribution, 44
- Boltzmann distribution, 166
- Boltzmann machine, 166
- Calculus of variations, 302
- CD, *see* contrastive divergence
- Centering trick (DBM), 312
- Central limit theorem, 45
- Chain rule of probability, 40
- Chess, 2
- Classical regularization, 128
- Classification, 71
- Cliffs, 151
- Clipping the gradient, 244
- Clique potential, *see* factor (graphical model)
- CNN, *see* convolutional neural network
- Collider, *see* explaining away
- Computer vision, 323
- Conditional computation, *see* dynamically structured nets, 318
- Conditional independence, 40
- Conditional probability, 39
- Constrained optimization, 65
- Context-specific independence, 169
- Contrast, 324
- Contrastive divergence, 284, 309, 310
- Convolution, 199, 313
- Convolutional neural network, 199
- Coordinate descent, 156, 310
- Correlation, 41
- Cost function, *see* objective function
- Covariance, 41
- Covariance matrix, 41
- curse of dimensionality, 100
- Cyc, 2
- D-separation, 169
- Dataset augmentation, 324, 329
- DBM, *see* deep Boltzmann machine
- Deep belief network, 296, 306, 307, 314
- Deep Blue, 2
- Deep Boltzmann machine, 296, 306, 308, 310, 314
- Deep learning, 2, 5
- Denoising score matching, 293
- Density estimation, 71
- density estimation, 92
- Derivative, 58
- Detector layer, 204
- Dirac delta function, 47
- Directed graphical model, 162
- Directional derivative, 62
- domain adaptation, 193
- Dot product, 23
- Doubly block circulant matrix, 201
- Dream sleep, 284, 304
- DropConnect, 149
- Dropout, 146, 310
- Dynamically structured networks, 318
- E-step, 299
- Early stopping, 114, 134, 136–138, 150

EBM, *see* energy-based model
 Effective number of parameters, 131
 Eigendecomposition, 28
 Eigenvalue, 29
 Eigenvector, 28
 ELBO, *see* evidence lower bound
 Element-wise product, *see* Hadamard product
 EM, *see* expectation maximization
 Empirical distribution, 47
 Energy function, 166
 Energy-based model, 166, 308
 Ensemble methods, 144
 Equality constraint, 67
 Equivariance, 202
 Error function, *see* objective function
 Euclidean norm, 26
 Euler-Lagrange equation, 302
 Evidence lower bound, 296, 298–300, 308
 Expectation, 41
 Expectation maximization, 298
 Expected value, *see* expectation
 Explaining away, 170

 Factor (graphical model), 164
 Factor graph, 169
 Factors of variation, 5
 Frequentist probability, 37
 Functional derivatives, 302

 Gaussian distribution, *see* Normal distribution 45
 Gaussian mixture, 48
 GCN, *see* Global contrast normalization
 Gibbs distribution, 165
 Gibbs sampling, 173
 Global contrast normalization, 325
 Global minimum, 13
 Gradient, 62
 Gradient clipping, 244
 Gradient descent, 62
 Graphical model, *see* structured probabilistic model

 Hadamard product, 22
 Harmonium, *see* Restricted Boltzmann machine 177
 Harmony theory, 167
 Helmholtz free energy, *see* evidence lower bound
 Hessian matrix, 63
 Identity matrix, 24

 Independence, 40
 Inequality constraint, 67
 Inference, 159, 174, 296, 298–300, 302, 304
 Invariance, 207

 Jacobian matrix, 52, 62
 Joint probability, 38

 Karush-Kuhn-Tucker conditions, 68
 Kernel (convolution), 200
 KKT conditions, *see* Karush-Kuhn-Tucker conditions
 KL divergence, *see* Kullback-Leibler divergence 42
 Kullback-Leibler divergence, 42

 Lagrange function, *see* Lagrangian
 Lagrange multipliers, 67, 303
 Lagrangian, 67
 Learner, 3
 Line search, 62
 Linear combination, 25
 Linear dependence, 26
 Local conditional probability distribution, 162
 Local minimum, 13
 Logistic regression, 3
 Logistic sigmoid, 48
 Loss function, *see* objective function

 M-step, 299
 Machine learning, 3
 Manifold hypothesis, 252
 manifold hypothesis, 100
 Manifold learning, 99, 252
 MAP inference, 300
 Marginal probability, 39
 Markov chain, 171
 Markov network, *see* undirected model 164
 Markov random field, *see* undirected model 164
 Matrix, 21
 Matrix inverse, 24
 Matrix product, 22
 Max pooling, 207
 Mean field, 309, 310
 Measure theory, 51
 Measure zero, 52
 Method of steepest descent, *see* gradient descent
 Missing inputs, 71
 Mixing (Markov chain), 175
 Mixture distribution, 48

MNIST, 310
 Model averaging, 144
 Moore-Penrose pseudoinverse, 139
 MP-DBM, *see* multi-prediction DBM
 Multi-modal learning, 197
 Multi-prediction DBM, 309, 312
 Multinomial distribution, 44
 Multinoulli distribution, 44

 Naive Bayes, 53
 Nat, 42
 Negative definite, 63
 Negative phase, 283
 Netflix Grand Prize, 146
 Noise-contrastive estimation, 293
 Norm, 26
 Normal distribution, 45, 47
 Normal equations, 131
 Normalized probability distribution, 165

 Object detection, 323
 Object recognition, 323
 Objective function, 12, 58
 one-shot learning, 196
 Orthogonality, 28
 Overfitting, 79

 Parameter sharing, 202
 Partial derivative, 58
 Partition function, 103, 165, 276, 309
 PCA, *see* principal components analysis
 PCD, *see* stochastic maximum likelihood
 Persistent contrastive divergence, *see* stochastic maximum likelihood
 Pooling, 199, 313
 Positive definite, 63
 Positive phase, 283
 Precision (of a normal distribution), 45, 47
 Predictive sparse decomposition, 183, 265
 Preprocessing, 324
 Principal components analysis, 31, 296, 326
 Principle components analysis, 90
 Probabilistic max pooling, 313
 Probability density function, 38
 Probability distribution, 37
 Probability mass function, 37
 Product rule of probability, *see* chain rule of probability
 PSD, *see* predictive sparse decomposition
 Pseudolikelihood, 289

 Random variable, 37
 Ratio matching, 292
 RBM, *see* restricted Boltzmann machine
 Receptive field, 203
 Regression, 71
 Regularization, 127
 Representation learning, 3
 Restricted Boltzmann machine, 177, 192, 296, 305, 306, 310, 312, 313
 Ridge regression, 129

 Scalar, 20
 Score matching, 291
 Second derivative, 62
 Second derivative test, 63
 Self-information, 42
 Separable convolution, 216
 Separation (probabilistic modeling), 167
 Shannon entropy, 42, 303
 Sigmoid, *see* logistic sigmoid
 Singular value decomposition, 30, 140
 SML, *see* stochastic maximum likelihood
 Softmax, 110
 Softplus, 48
 Spam detection, 3
 Sparse coding, 191, 296
 spectral radius, 236
 Sphering, *see* Whitening, 326
 Square matrix, 26
 Standard deviation, 41
 Statistic, 83
 Steepest descent, *see* gradient descent
 Stochastic gradient descent, 310
 Stochastic maximum likelihood, 288, 309, 310
 Stochastic pooling, 149
 Structure learning, 173
 Structured output, 71
 Structured probabilistic model, 158
 Sum rule of probability, 39
 SVD, *see* singular value decomposition
 Symmetric matrix, 28

 Tangent plane, 255
 Tensor, 21
 Test example, 12
 Tiled convolution, 212
 Toeplitz matrix, 201
 Trace operator, 31
 Training criterion, 12
 Transcription, 71

- Transfer learning, 193
- Transpose, 21
- Triangle inequality, 26
- Unbiased, 84
- Underfitting, 78
- Undirected model, 164
- Uniform distribution, 38
- Unit norm, 28
- Unnormalized probability distribution, 164
- V-structure, *see* explaining away
- Variance, 41
- Variational derivatives, *see* functional derivatives
- Variational free energy, *see* evidence lower bound
- Vector, 20
- Weight decay, 129
- Whitening, 326
- ZCA, *see* zero-phase components analysis
- zero-data learning, 196
- Zero-phase components analysis, 326
- zero-shot learning, 196