References

Abel, Niels H. 1826. *Démonstration de l'Impossibilité de la Résolution Algébrique des Équations Générales qui Passent le Quatrième Degré*. Grøndahl & Søn. pages 336

Adhikari, Ani, and DeNero, John. 2018. *Computational and Inferential Thinking: The Foundations of Data Science*. Gitbooks. pages 256

Agakov, Felix V., and Barber, David. 2004. An Auxiliary Variational Method. In: *Neural Information Processing Systems*. pages 280

Agarwal, Arvind, and III, Hal Daumé. 2010. A Geometric View of Conjugate Priors. *Machine Learning*, **81**(1), 99–113. pages 212

Agresti, A. 2002. *Categorical Data Analysis*. Wiley. pages 274

Akaike, Hirotugu. 1974. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723. pages 290

Akhiezer, N.I., and Glazman, I.M. 1993. *Theory of Linear Operators in Hilbert Space*. Dover Publications, Inc. pages 397

Alpaydin, Ethem. 2010. *Introduction to Machine Learning*. MIT Press. pages 2

Amari, Shun-ichi. 2016. *Information Geometry and Its Applications*. Springer. pages 201

Argyriou, Andreas, and Dinuzzo, Francesco. 2014. A Unifying View of Representer Theorems. In: *Proceedings of the International Conference on Machine Learning*. pages 387

Aronszajn, N. 1950. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, **68**, 337–404. pages 392, 393, 397

Axler, Sheldon. 2015. *Linear Algebra Done Right*. 3rd edn. Springer. pages 51, 62, 65, 95

Bakir, Gökhan, Hofmann, Thomas, Schölkopf, Bernhard, Smola, Alexander J., Taskar, Ben, and Vishwanathan, S.V.N (eds). 2007. *Predicting Structured Data*. MIT Press. pages 285

Barber, David. 2012. *Bayesian Reasoning and Machine Learning*. Cambridge University Press. pages 2, 226, 276, 279, 317, 370

Barndorff-Nielsen, Ole. 2014. *Information and Exponential Families: In Statistical Theory*. Wiley. pages 225

Bartholomew, David, Knott, Martin, and Moustaki, Irini. 2011. *Latent Variable Models and Factor Analysis: A Unified Approach*. Wiley. pages 348

Beck, Amir, and Teboulle, Marc. 2003. Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization. *Operations Research Letters*, **31**(3), 167–175. pages 250

Belabbas, Mohamed-Ali, and Wolfe, Patrick J. 2009. Spectral Methods in Machine Learning and New Strategies for Very Large Datasets. *Proceedings of the National Academy of Sciences*, 0810600105. pages 138

Belkin, Mikhail, and Niyogi, Partha. 2003. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, **15**(6), 1373–1396. pages 138

Draft chapter (December 21, 2018) from "Mathematics for Machine Learning" ©2018 by Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong. To be published by Cambridge University Press. Report errata and feedback to `http://mml-book.com`. Please do not post or distribute this file, please link to `https://mml-book.com`.

Ben-Hur, Asa, Ong, Cheng S., Sonnenburg, Sören, Schölkopf, Bernhard, and Rätsch, Gunnar. 2008. Support Vector Machines and Kernels for Computational Biology. *PLoS Computational Biology*, **4**(10), e1000173. pages 393, 397

Bennett, Kristin P., and Bredensteiner, Erin J. 2000a. Duality and Geometry in SVM Classifiers. In: *Proceedings of the International Conference on Machine Learning*. pages 391

Bennett, Kristin P., and Bredensteiner, Erin J. 2000b. Geometry in Learning. In: Gorini, Catherine A. (ed), *Geometry at Work*. The Mathematical Association of America. pages 391

Berlinet, Alain, and Thomas-Agnan, Christine. 2004. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer. pages 392

Bertsekas, Dimitri P. 1999. *Nonlinear Programming*. Athena Scientific. pages 250

Bertsekas, Dimitri P. 2009. *Convex Optimization Theory*. Athena Scientific. pages 251

Betancourt, Michael. 2018. *Probability Theory (for Scientists and Engineers)*. `https://betanalpha.github.io/assets/case_studies/probability_theory.html`. pages 184

Bickel, Peter J., and Doksum, Kjell. 2006. *Mathematical Statistics, Basic Ideas and Selected Topics*. Vol. 1. Prentice Hall. pages 225

Bickson, Danny, Dolev, Danny, Shental, Ori, Siegel, Paul H., and Wolf, Jack K. 2007. Linear Detection via Belief Propagation. In: *Proceedings of the Annual Allerton Conference on Communication, Control, and Computing*. pages 285

Billingsley, Patrick. 1995. *Probability and Measure*. Wiley. pages 184, 223, 225

Bishop, Christopher M. 1995. *Neural Networks for Pattern Recognition*. Clarendon Press. pages 317

Bishop, Christopher M. 1999. Bayesian PCA. In: *Advances in Neural Information Processing Systems*. pages 348

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Springer. pages vii, 2, 96, 173, 177, 182, 196, 203, 213, 217, 221, 226, 266, 276, 277, 278, 279, 280, 284, 302, 317, 335, 369, 370, 397, 398

Blei, David M., Kucukelbir, Alp, and McAuliffe, Jon D. 2017. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, **112**(518), 859–877. pages 276, 279, 348

Blum, Arvim, and Hardt, Moritz. 2015. The Ladder: A Reliable Leaderboard for Machine Learning Competitions. In: *International Conference on Machine Learning*. pages 265

Bonnans, J. Frédéric, Gilbert, J. Charles, Lemaréchal, Claude, and Sagastizábal, Claudia A. 2006. *Numerical Optimization: Theoretical and Practical Aspects*. 2nd edn. Springer Verlag. pages 251

Borwein, Jonathan M., and Lewis, Adrian S. 2006. *Convex Analysis and Nonlinear Optimization*. 2nd edn. Canadian Mathematical Society. pages 251

Bottou, Léon. 1998. Online Algorithms and Stochastic Approximations. In: *Online Learning and Neural Networks*. Cambridge University Press. pages 236

Bottou, Léon, Curtis, Frank E, and Nocedal, Jorge. 2018. Optimization Methods for Large-scale Machine Learning. *SIAM Review*, **60**(2), 223–311. pages 237, 250, 251

Boucheron, Stephane, Lugosi, Gabor, and Massart, Pascal. 2013. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press. pages 181

Boyd, Stephen, and Vandenberghe, Lieven. 2004. *Convex Optimization*. Cambridge University Press. pages 234, 238, 241, 251

Boyd, Stephen, and Vandenberghe, Lieven. 2018. *Introduction to Applied Linear Algebra*. Cambridge University Press. pages 95

Brochu, Eric, Cora, Vlad M., and de Freitas, Nando. 2009. *A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning*. Tech. rept. TR-2009-023. Department of Computer Science, University of British Columbia. pages 276

Brooks, Steve, Gelman, Andrew, Jones, Galin L., and Meng, Xiao-Li (eds). 2011. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC. pages 279

Brown, Lawrence D. 1986. *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. Lecture Notes - Monograph Series. Institute of Mathematical Statistics. pages 215, 217

Bryson, Arthur E. 1961. A Gradient Method for Optimizing Multi-stage Allocation Processes. In: *Proceedings of the Harvard University Symposium on Digital Computers and Their Applications*. pages 162

Bubeck, Sébastien. 2015. Convex Optimization: Algorithms and Complexity. *Foundations and Trends in Machine Learning*, **8**(3-4), 231–357. pages 251

Bühlmann, Peter, and Geer, Sara Van De. 2011. *Statistics for High-Dimensional Data*. Springer. pages 268

Burges, Christopher. 2010. Dimension Reduction: A Guided Tour. *Foundations and Trends in Machine Learning*, **2**(4), 275–365. pages 349

Carroll, J Douglas, and Chang, Jih-Jie. 1970. Analysis of Individual Differences in Multidimensional Scaling via an N-way Generalization of "Eckart-Young" Decomposition. *Psychometrika*, **35**(3), 283–319. pages 138, 139

Casella, George, and Berger, Roger L. 2002. *Statistical Inference*. Duxbury. pages 191, 219, 220, 274

Çinlar, Erhan. 2011. *Probability and Stochastics*. Springer. pages 225

Chang, Chih-Chung, and Lin, Chih-Jen. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. pages 396

Cheeseman, Peter. 1985. In Defense of Probability. In: *Proceedings of the International Joint Conference on Artificial Intelligence*. pages 274

Chollet, Francois, and Allaire, J. J. 2018. *Deep Learning with R*. Manning Publications. pages 2

Codd, Edgar F. 1990. *The Relational Model for Database Management*. Addison-Wesley Longman Publishing. pages 256

Cunningham, John P., and Ghahramani, Zoubin. 2015. Linear Dimensionality Reduction: Survey, Insights, and Generalizations. *Journal of Machine Learning Research*, **16**, 2859–2900. pages 349

Datta, Biswa N. 2010. *Numerical Linear Algebra and Applications*. Vol. 116. SIAM. pages 130

Davidson, Anthony C., and Hinkley, David V. 1997. *Bootstrap Methods and their Application*. Cambridge University Press. pages 268

Dean, Jeffrey, Corrado, Greg S, Monga, Rajat, Chen, Kai, Devin, Matthieu, Le, Quoc V, Mao, Mark Z, Ranzato, Marc Aurelio, Senior, Andrew, Tucker, Paul, Yang, Ke, and Ng, Andrew Y. 2012. Large Scale Distributed Deep Networks. In: *Advances in Neural Information Processing Systems*. pages 237

Deisenroth, Marc P., and Mohamed, Shakir. 2012. Expectation Propagation in Gaussian Process Dynamical Systems. Pages 2618–2626 of: *Advances in Neural Information Processing Systems*. pages 285

Deisenroth, Marc P., and Ohlsson, Henrik. 2011. A General Perspective on Gaussian Filtering and Smoothing: Explaining Current and Deriving New Algorithms. In: *Proceedings of the American Control Conference*. pages 203

Deisenroth, Marc P., Fox, Dieter, and Rasmussen, Carl E. 2015. Gaussian Processes for Data-Efficient Learning in Robotics and Control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**(2), 408–423. pages 92, 190, 276

©2018 Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong. To be published by Cambridge University Press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, **39**(1), 1–38. pages 363

Deng, Li, Seltzer, Michael L., Yu, Dong, Acero, Alex, Mohamed, Abdel-rahman, and Hinton, Geoffrey E. 2010. Binary Coding of Speech Spectrograms using a Deep Auto-Encoder. Pages 1692–1695 of: *Interspeech*. pages 84

Devroye, Luc. 1986. *Non-Uniform Random Variate Generation*. Springer. pages 208

Donoho, David L, and Grimes, Carrie. 2003. Hessian Eigenmaps: Locally Linear Embedding Techniques for High-dimensional Data. *Proceedings of the National Academy of Sciences*, **100**(10), 5591–5596. pages 138

Dostál, Zdeněk. 2009. *Optimal Quadratic Programming Algorithms: With Applications to Variational Inequalities*. Springer. pages 388

Douven, Igor. 2017. Abduction. In: *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. pages 261

Downey, Allen B. 2014. *Think Stats: Exploratory Data Analysis*. 2nd edn. O'Reilly Media. pages 225

Dreyfus, Stuart. 1962. The Numerical Solution of Variational Problems. *Journal of Mathematical Analysis and Applications*, **5**(1), 30–45. pages 162

Drumm, Volker, and Weil, Wolfgang. 2001. *Lineare Algebra und Analytische Geometrie*. Lecture Notes, Universität Karlsruhe (TH). pages 20, 56

Dudley, Richard M. 2002. *Real Analysis and Probability*. Cambridge University Press. pages 225

Eaton, Morris L. 2007. *Multivariate Statistics: A Vector Space Approach*. Vol. 53. Institute of Mathematical Statistics Lecture Notes—Monograph Series. pages 199

Efron, Bradley, and Hastie, Trevor. 2016. *Computer Age Statistical Inference: Algorithms, Evidence and Data Science*. Cambridge University Press. pages 177, 209, 273, 274

Efron, Bradley, and Tibshirani, Robert J. 1993. *An Introduction to the Bootstrap*. Chapman and Hall/CRC. pages 268

Elliott, Conal. 2009. Beautiful Differentiation. In: *International Conference on Functional Programming*. pages 173

Evgeniou, Theodoros, Pontil, Massimiliano, and Poggio, Tomaso. 2000. Statistical Learning Theory: A Primer. *International Journal of Computer Vision*, **38**(1), 9–13. pages 268

Fan, Rong-En, Chang, Kai-Wei, Hsieh, Cho-Jui, Wang, Xiang-Rui, and Lin, Chih-Jen. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, **9**, 1871–1874. pages 383

Gal, Yarin, van der Wilk, Mark, and Rasmussen, Carl E. 2014. Distributed Variational Inference in Sparse Gaussian Process Regression and Latent Variable Models. In: *Advances in Neural Information Processing Systems*. pages 237

Gärtner, Thomas. 2008. *Kernels for Structured Data*. World Scientific. pages 393

Gavish, Matan, and Donoho, David L. 2014. The Optimal Hard Threshold for Singular Values is $4\sqrt{3}$. *IEEE Transactions on Information Theory*, **60**(8), 5040–5053. pages 347

Gelman, Andrew, Carlin, John B., Stern, Hal S., and Rubin, Donald B. 2004. *Bayesian Data Analysis*. Second. Chapman & Hall/CRC. pages 275

Gentle, James E. 2004. *Random Number Generation and Monte Carlo Methods*. 2nd edn. Springer. pages 208

Ghahramani, Zoubin. 2015. Probabilistic Machine Learning and Artificial Intelligence. *Nature*, **521**, 452–459. pages 188, 279

Ghahramani, Zoubin, and Roweis, Sam T. 1999. Learning Nonlinear Dynamical Systems using an EM Algorithm. In: *Advances in Neural Information Processing Systems*. MIT Press. pages 277, 370

Gilks, Walter R., Richardson, Sylvia, and Spiegelhalter, David J. 1996. *Markov Chain Monte Carlo in Practice*. Chapman & Hall. pages 276, 279, 348

Gneiting, Tilmann, and Raftery, Adrian E. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, **102**(477), 359–378. pages 398

Goh, Gabriel. 2017. Why Momentum Really Works. *Distill*. pages 234, 250

Gohberg, Israel, Goldberg, Seymour, and Krupnik, Nahum. 2012. *Traces and Determinants of Linear Operators*. Vol. 116. Birkhäuser. pages 104

Golan, Jonathan S. 2007. *The Linear Algebra a Beginning Graduate Student Ought to Know*. 2nd edn. Springer. pages 65

Golub, Gene H., and Van Loan, Charles F. 2012. *Matrix Computations*. Vol. 4. JHU Press. pages 65

Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron. 2016. *Deep Learning*. MIT Press. pages 217, 225, 258, 285, 317

Graepel, Thore, Candela, Joaquin Quiñonero-Candela, Borchert, Thomas, and Herbrich, Ralf. 2010. Web-scale Bayesian Click-through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine. In: *Proceedings of the International Conference on Machine Learning*. pages 276

Griewank, Andreas, and Walther, Andrea. 2003. Introduction to Automatic Differentiation. In: *Proceedings in Applied Mathematics and Mechanics*. pages 173

Griewank, Andreas, and Walther, Andrea. 2008. *Evaluating Derivatives, Principles and Techniques of Algorithmic Differentiation*. second edn. SIAM, Philadelphia. pages 173

Grimmett, Geoffrey, and Welsh, Dominic. 2014. *Probability: an Introduction*. 2nd edn. Oxford University Press. pages 225

Grinstead, Charles M., and Snell, J. Laurie. 1997. *Introduction to Probability*. American Mathematical Society. pages 177, 201, 225

Hacking, Ian. 2001. *Probability and Inductive Logic*. Cambridge University Press. pages 225

Hall, Peter. 1992. *The Bootstrap and Edgeworth Expansion*. Springer. pages 268

Hallin, Marc, Paindaveine, Davy, and Šiman, Miroslav. 2010. Multivariate quantiles and multiple-output regression quantiles: from ℓ_1 optimization to halfspace depth. *Annals of Statistics*, **38**, 635–669. pages 192

Hasselblatt, Boris, and Katok, Anatole. 2003. *A first course in dynamics with a Panorama of Recent Developments*. Cambridge University Press. pages 179

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. 2001. *The Elements of Statistical Learning—Data Mining, Inference, and Prediction*. Springer Series in Statistics. 175 Fifth Avenue, New York City, NY, USA: Springer-Verlag New York, Inc. pages 268

Hazan, Elad. 2015. Introduction to Online Convex Optimization. *Foundations and Trends in Optimization*, **2**(3-4), 157–325. pages 251

Hensman, James, Fusi, Nicolò, and Lawrence, Neil D. 2013. Gaussian Processes for Big Data. In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. pages 237

Herbrich, Ralf, Minka, Tom, and Graepel, Thore. 2007. TrueSkill(TM): A Bayesian Skill Rating System. In: *Advances in Neural Information Processing Systems*. pages 276, 285

Hiriart-Urruty, Jean-Baptiste, and Lemaréchal, Claude. 2001. *Fundamentals of Convex Analysis*. Springer. pages 247, 251

Hoffman, Matthew D., Blei, David M., Wang, Chong, and Paisley, John. 2013. Stochastic Variational Inference. *Journal of Machine Learning Research*, **14**(1), 1303–1347. pages 237, 276

©2018 Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong. To be published by Cambridge University Press.

Hofmann, Thomas, Schölkopf, Bernhard, and Smola, Alexander J. 2008. Kernel Methods in Machine Learning. *Annals of Statistics*, **36**(3), 1171–1220. pages 387

Hogben, Leslie. 2013. *Handbook of Linear Algebra*. 2nd edn. Chapman and Hall/CRC. pages 20, 108

Horn, Roger A., and Johnson, Charles R. 2013. *Matrix Analysis*. Cambridge University Press. pages 65

Hotelling, Harold. 1933. Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, **24**, 417–441. pages 83, 96, 319, 322

Hyvarinen, Aapo, Oja, Erkki, and Karhunen, Juha. 2001. *Independent Component Analysis*. Wiley. pages 348

Imbens, Guido W., and Rubin, Donald B. 2015. *Causal Inference for Statistics, Social and Biomedical Sciences*. Cambridge University Press. pages 285

Jacod, Jean, and Protter, Philip. 2004. *Probability Essentials*. 2nd edn. Springer. pages 178, 180, 184, 225

Jaynes, Edwin T. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press. pages 175, 176, 177, 187, 225, 274

Jefferys, Willian H., and Berger, James O. 1992. Ockham's Razor and Bayesian Analysis. *American Scientist*, **80**, 64–72. pages 287

Jeffreys, Harold. 1961. *Theory of Probability*. 3rd edn. Oxford University Press. pages 289

Joachims, Thorsten. 1999. *Advances in Kernel Methods—Support Vector Learning*. MIT Press. Chap. Making Large-Scale SVM Learning Practical, pages 169–184. pages 396

Jordan, Michael I., Ghahramani, Zoubin, Jaakkola, Tommi S., and Saul, Lawrence K. 1999. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, **37**, 183–233. pages 276, 279

Julier, Simon J., and Uhlmann, Jeffrey K. 1997. A New Extension of the Kalman Filter to Nonlinear Systems. In: *Proceedings of AeroSense Symposium on Aerospace/Defense Sensing, Simulation and Controls*. pages 173

Kalman, Dan. 1996. A Singularly Valuable Decomposition: The SVD of a Matrix. *The College Mathematics Journal*, **27**(1), 2–23. pages 120

Kalman, Rudolf E.f. 1960. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME—Journal of Basic Engineering*, **82**(Series D), 35–45. pages 203

Kamthe, Sanket, and Deisenroth, Marc P. 2018. Data-Efficient Reinforcement Learning with Probabilistic Model Predictive Control. In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*. pages 276

Katz, Victor J. 2004. *A History of Mathematics*. Pearson/Addison-Wesley. pages 110

Kelley, Henry J. 1960. Gradient Theory of Optimal Flight Paths. *Ars Journal*, **30**(10), 947–954. pages 162

Kimeldorf, George S., and Wahba, Grace. 1970. A Correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines. *Annals of Mathematical Statistics*, **41**(2), 495–502. pages 387

Kittler, J., and Föglein, J. 1984. Contextual Classification of Multispectral Pixel Data. *Image and Vision Computing*, **2**(1), 13–29. pages 285

Kolda, Tamara G., and Bader, Brett W. 2009. Tensor Decompositions and Applications. *SIAM Review*, **51**(3), 455–500. pages 139

Koller, Daphne, and Friedman, Nir. 2009. *Probabilistic Graphical Models*. MIT Press. pages 284

Kong, Linglong, and Mizera, Ivan. 2012. Quantile Tomography: Using Quantiles with Multivariate Data. *Statistica Sinica*, **22**, 1598–1610. pages 192

Lang, Serge. 1987. *Linear Algebra*. Springer. pages 116

Lawrence, Neil. 2005. Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models. *Journal of Machine Learning Research*, **6**(Nov.), 1783–1816. pages 349

Leemis, Lawrence M., and McQueston, Jacquelyn T. 2008. Univariate Distribution Relationships. *The American Statistician*, **62**(1), 45–53. pages 208, 211

Lehmann, Erich L., and Romano, Joseph P. 2005. *Testing Statistical Hypotheses*. Springer. pages 221

Lehmann, Erich Leo, and Casella, George. 1998. *Theory of Point Estimation*. Springer. pages 214, 215, 225, 273

Liesen, Jörg, and Mehrmann, Volker. 2015. *Linear Algebra*. Springer. pages 20, 37, 65

Ljung, Lennart. 1999. *System Identification: Theory for the User*. Prentice Hall. pages 277

Loosli, Gaëlle, Canu, Stéphane, and Ong, Cheng S. 2016. Learning SVM in Kreĭn Spaces. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, **38**(6), 1204–1216. pages 397

Luenberger, David G. 1969. *Optimization by Vector Space Methods*. Wiley. pages 251

Maaløe, Lars, Sønderby, Casper Kaae, Søren K, Sønderby, and Winther, Ole. 2016. Auxiliary Deep Generative Models. In: *Proceedings of the International Conference on Machine Learning*. pages 280

MacKay, Davic J. C. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press. pages 2, 173, 189, 225, 226, 287, 346, 370

MacKay, David J. C. 1992. Bayesian Interpolation. *Neural Computation*, **4**, 415–447. pages 287

MacKay, David J. C. 1998. Introduction to Gaussian Processes. Pages 133–165 of: *Neural Networks and Machine Learning*, vol. 168. Berlin, Germany: Springer. pages 318

Magnus, Jan R., and Neudecker, Heinz. 2007. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. 3rd edn. Wiley. pages 173

Manton, Jonathan H., and Amblard, Pierre-Olivier. 2015. A Primer on Reproducing Kernel Hilbert Spaces. *Foundations and Trends in Signal Processing*, **8**(1–2), 1–126. pages 397

Markovsky, Ivan. 2011. *Low Rank Approximation: Algorithms, Implementation, Applications*. Springer. pages 139

Maybeck, Peter S. 1979. *Stochastic Models, Estimation, and Control*. Mathematics in Science and Engineering, vol. 141. Academic Press, Inc. pages 173, 277

McCullagh, Peter, and Nelder, John A. 1989. *Generalized Linear Models*. 2nd edn. CRC Press. pages 274

McEliece, Robert J., MacKay, David J. C., and Cheng, Jung-Fu. 1998. Turbo Decoding as an Instance of Pearl's "Belief Propagation" Algorithm. *IEEE Journal on Selected Areas in Communications*, **16**(2), 140–152. pages 285

Meyer, Carl D. 2000. *Matrix Analysis and Applied Linear Algebra*. Vol. 71. SIAM. pages 108

Mika, Sebastian, Rätsch, Gunnar, Weston, Jason, Schölkopf, Bernhard, and Müller, Klaus-Robert. 1999. Fisher Discriminant Analysis with Kernels. *Neural Networks for Signal Processing*, **IX**, 41–48. pages 138

Minka, Thomas P. 2001a. *A Family of Algorithms for Approximate Bayesian Inference*. Ph.D. thesis, Massachusetts Institute of Technology. pages 276

Minka, Tom. 2001b. Automatic Choice of Dimensionality of PCA. In: *Advances in Neural Information Processing Systems*. pages 348

Mitchell, Tom. 1997. *Machine Learning*. McGraw Hill. pages 13, 266

©2018 Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong. To be published by Cambridge University Press.

Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Rusu, Andrei A., Veness, Joel, Bellemare, Marc G., Graves, Alex, Riedmiller, Martin, Fidjeland, Andreas K., Ostrovski, Georg, Petersen, Stig, Beattie, Charles, Sadik, Amir, Antonoglou, Ioannis, King, Helen, Kumaran, Dharshan, Wierstra, Daan, Legg, Shane, and Hassabis, Demis. 2015. Human-Level Control through Deep Reinforcement Learning. *Nature*, **518**, 529–533. pages 237

Moonen, Marc, and De Moor, Bart. 1995. *SVD and Signal Processing, III: Algorithms, Architectures and Applications*. Elsevier. pages 139

Moustaki, Irini., Knott, Martin., and Bartholomew, David. J. 2015. *Latent-Variable Modeling*. American Cancer Society. Pages 1–10. pages 277, 278, 279

Müller, Andreas C., and Guido, Sarah. 2016. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Publishing. pages 2

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press. pages 2, 173, 203, 211, 226, 276, 277, 278, 279, 287, 289, 317, 335, 348, 365, 370, 397

Neal, Radford M. 1996. *Bayesian Learning for Neural Networks*. Ph.D. thesis, Department of Computer Science, University of Toronto. pages 318

Neal, Radford M., and Hinton, Geoffrey E. 1999. A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants. Pages 355–368 of: *Learning in Graphical Models*. MIT Press. pages 363

Nelsen, Roger. 2006. *An Introduction to Copulas*. Springer. pages 221

Nesterov, Yuri. 2018. *Lectures on Convex Optimization*. Springer. pages 250, 251

Neumaier, Arnold. 1998. Solving Ill-Conditioned and Singular Linear Systems: A Tutorial on Regularization. *SIAM Review*, **40**, 636–666. pages 268

Nocedal, Jorge, and Wright, Stephen J. 2006. *Numerical Optimization*. Springer. pages 167, 250, 396

Nowozin, Sebastian, Gehler, Peter V., Jancsary, Jeremy, and Lampert, Christoph H. (eds). 2014. *Advanced Structured Prediction*. MIT Press. pages 285

O'Hagan, Anthony. 1991. Bayes-Hermite Quadrature. *Journal of Statistical Planning and Inference*, **29**, 245–260. pages 289

Ong, Cheng S., Mary, Xavier, Canu, Stéphane, and Smola, Alexander J. 2004. Learning with Non-positive Kernels. Pages 639–646 of: *Proceedings of the International Conference on Machine Learning*. pages 397

Ormoneit, Dirk, Sidenbladh, Hedvig, Black, Michael J., and Hastie, Trevor. 2001. Learning and Tracking Cyclic Human Motion. In: *Advances in Neural Information Processing Systems*. pages 139

Page, Lawrence, Brin, Sergey, Motwani, Rajeev, and Winograd, Terry. 1999. *The PageRank Citation Ranking: Bringing Order to the Web*. Tech. rept. Stanford InfoLab. pages 113, 336

Paquet, Ulrich. 2008. *Bayesian Inference for Latent Variable Models*. Ph.D. thesis, University of Cambridge. pages 277, 278, 279

Parzen, Emanuel. 1962. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, **33**(3), 1065–1076. pages 371

Pearl, Judea. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann. pages 176, 283

Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference*. 2nd edn. Cambridge University Press. pages 281, 285

Pearson, Karl. 1895. Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **186**, 343–414. pages 371

Pearson, Karl. 1901. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, **2**(11), 559–572. pages 83, 96, 138, 319, 329

Peters, Jonas, Janzing, Dominik, and Schölkopf, Bernhard. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press. pages 285

Petersen, K. B., and Pedersen, M. S. 2012 (nov). *The Matrix Cookbook*. Tech. rept. Technical University of Denmark. Version 20121115. pages 161

Pollard, David. 2002. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press. pages 225

Polyak, Roman A. 2016. The Legendre Transformation in Modern Optimization. Pages 437–507 of: *Optimization and Its Applications in Control and Data Sciences*. Springer. pages 251

Press, William H., Teukolsky, Saul A., Vetterling, William T., and Flannery, Brian P. 2007. *Numerical Recipes: The Art of Scientific Computing*. 3rd edn. Cambridge University Press. pages 130, 138

Proschan, Michael A., and Presnell, Brett. 1998. Expect the Unexpected from Conditional Expectation. *American Statistician*, **52**(3), 248–252. pages 225

Ranganath, Rajesh, Tran, Dustin, and Blei, David M. 2016. Hierarchical Variational Models. In: *Proceedings of the International Conference on Machine Learning*. pages 280

Raschka, Sebastian, and Mirjalili, Vahid. 2017. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*. Packt Publishing. pages 2

Rasmussen, Carl E., and Ghahramani, Zoubin. 2001. Occam's Razor. In: *Advances in Neural Information Processing Systems*. pages 290

Rasmussen, Carl E., and Ghahramani, Zoubin. 2003. Bayesian Monte Carlo. In: *Advances in Neural Information Processing Systems*. pages 289

Rasmussen, Carl E., and Williams, Christopher K. I. 2006. *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press. pages 95, 203, 226, 318, 393

Reid, Mark, and Williamson, Robert C. 2011. Information, Divergence and Risk for Binary Experiments. *Journal of Machine Learning Research*, **12**, 731–817. pages 398

Rezende, Danilo J., and Mohamed, Shakir. 2015. Variational Inference with Normalizing Flows. In: *Proceedings of the International Conference on Machine Learning*. pages 225

Rifkin, Ryan M., and Lippert, Ross A. 2007. Value Regularization and Fenchel Duality. *Journal of Machine Learning Research*, **8**, 441–479. pages 397

Rockafellar, R. Tyrrell. 1970. *Convex Analysis*. Princeton University Press. pages 251

Rogers, Simon, and Girolami, Mark. 2016. *A First Course in Machine Learning*. 2nd edn. Chapman and Hall/CRC. pages 2, 369

Rosenbaum, Paul R. 2017. *Observation & Experiment: An Introduction to Causal Inference*. Harvard University Press. pages 285

Rosenblatt, Murray. 1956. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, **27**(3), 832–837. pages 371

Roweis, Sam, and Ghahramani, Zoubin. 1999. A Unifying Review of Linear Gaussian Models. *Neural Computation*, **11**(2), 305–345. pages 203, 370

Roweis, Sam T. 1998. EM Algorithms for PCA and SPCA. Pages 626–632 of: *Advances in Neural Information Processing Systems*. pages 347

Roy, Anindya, and Banerjee, Sudipto. 2014. *Linear Algebra and Matrix Analysis for Statistics*. Chapman and Hall/CRC. pages 120

Rubinstein, Reuven Y., and Kroese, Dirk P. 2016. *Simulation and the Monte Carlo Method*. Vol. 10. Wiley. pages 138

Ruffini, Paolo. 1799. *Teoria Generale delle Equazioni, in cui si Dimostra Impossibile la Soluzione Algebraica delle Equazioni Generali di Grado Superiore al Quarto*. Stamperia di S. Tommaso d'Aquino. pages 336

©2018 Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong. To be published by Cambridge University Press.

Rumelhart, David E., Hinton, Geoffrey E., and Williams, Ronald J. 1986. Learning Representations by Back-propagating Errors. *Nature*, **323**(6088), 533–536. pages 162, 234

Sæmundsson, Steindór, Hofmann, Katja, and Deisenroth, Marc P. 2018. Meta Reinforcement Learning with Latent Variable Gaussian Processes. In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. pages 277

Saitoh, Saburou. 1988. *Theory of Reproducing Kernels and its Applications*. Longman Scientific & Technical. pages 393, 397

Salimans, Tim, Kingma, Diederik P., and Welling, Max. 2015. Markov Chain Monte Carlo and Variational Inference: Bridging the Gap. In: *Proceedings of the International Conference on Machine Learning*. pages 280

Schölkopf, Bernhard, and Smola, Alexander J. 2002. *Learning with Kernels—Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press. pages 95, 318, 349, 378, 389, 392, 393, 397

Schölkopf, Bernhard, Smola, Alexander J., and Müller, Klaus-Robert. 1997. Kernel Principal Component Analysis. In: *Proceedings of the International Conference on Artificial Neural Networks*. Springer. pages 95

Schölkopf, Bernhard, Smola, Alexander J., and Müller, Klaus-Robert. 1998. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, **10**(5), 1299–1319. pages 349

Schölkopf, Bernhard, Herbrich, Ralf, and Smola, Alexander J. 2001. A Generalized Representer Theorem. In: *Proceedings of the International Conference on Computational Learning Theory*. pages 387

Schwartz, Laurent. 1964. Sous espaces Hilbertiens d'espaces vectoriels topologiques et noyaux associés. *Journal d'Analyse Mathématique*, **13**, 115–256. in French. pages 397

Schwarz, Gideon E. 1978. Estimating the Dimension of a Model. *Annals of Statistics*, **6**(2), 461–464. pages 290

Shahriari, Bobak, Swersky, Kevin, Wang, Ziyu, Adams, Ryan P., and De Freitas, Nando. 2016. Taking the Human out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, **104**(1), 148–175. pages 276

Shalev-Shwartz, Shai, and Ben-David, Shai. 2014. *Understanding Machine Leanring: From Theory to Algorithms*. Cambridge University Press. pages 2, 181, 268, 378

Shawe-Taylor, John, and Cristianini, Nello. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press. pages 392, 397

Shawe-Taylor, John, and Sun, Shiliang. 2011. A Review of Optimization Methodologies in Support Vector Machines. *Neurocomputing*, **74**(17), 3609–3618. pages 396

Shental, O., Bickson, D., P. H. Siegel and, J. K. Wolf, and Dolev, D. 2008. Gaussian Belief Propagatio Solver for Systems of Linear Equations. In: *Proceedings of the International Symposium on Information Theory*. pages 285

Shewchuk, Jonathan R. 1994. *An Introduction to the Conjugate Gradient Method Without the Agonizing Pain*. pages 250

Shi, Jianbo, and Malik, Jitendra. 2000. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8), 888–905. pages 138

Shi, Qinfeng, Petterson, James, Dror, Gideon, Langford, John, Smola, Alex, and Vishwanathan, S.V.N. 2009. Hash Kernels for Structured Data. *Journal of Machine Learning Research*, 2615–2637. pages 393

Shiryayev, A. N. 1984. *Probability*. Springer. pages 225

Shor, Naum Z. 1985. *Minimization Methods for Non-differentiable Functions*. Springer. pages 250

Shotton, Jamie, Winn, John, Rother, Carsten, and Criminisi, Antonio. 2006. TextonBoost: Joint Appearance, Shape and Context Modeling for Mulit-Class Object Recognition and Segmentation. In: *Proceedings of the European Conference on Computer Vision*. pages 285

Smith, Adrian F. M., and Spiegelhalter, David. 1980. Bayes Factors and Choice Criteria for Linear Models. *Journal of the Royal Statistical Society B*, **42**(2), 213–220. pages 287

Snoek, Jasper, Larochelle, Hugo, and Adams, Ryan P. 2012. Practical Bayesian Optimization of Machine Learning Algorithms. In: *Advence in Neural Information Processing Systems*. pages 276

Spearman, Charles. 1904. "General Intelligence," Objectively Determined and Measured. *American Journal of Psychology*, **15**(2), 201–292. pages 348

Steinwart, Ingo. 2007. How to Compare Different Loss Functions and Their Risks. *Constructive Approximation*, **26**, 225–287. pages 397

Steinwart, Ingo, and Christmann, Andreas. 2008. *Support Vector Machines*. Springer. pages 372, 376, 378, 383, 397

Stoer, Josef, and Burlirsch, Roland. 2002. *Introduction to Numerical Analysis*. Springer. pages 37, 65, 95, 289

Strang, Gilbert. 1993. The Fundamental Theorem of Linear Algebra. *The American Mathematical Monthly*, **100**(9), 848–855. pages 120

Strang, Gilbert. 2003. *Introduction to Linear Algebra*. 3rd edn. Wellesley-Cambridge Press. pages 20, 37, 65, 81, 316

Stray, Jonathan. 2016. *The Curious Journalist's Guide to Data*. Tow Center for Digital Journalism at Columbia's Graduate School of Journalism. pages 256

Strogatz, Steven. 2014. Writing about Math for the Perplexed and the Traumatized. *Notices of the American Mathematical Society*, **61**(3), 286–291. pages 2

Sucar, Luis E., and Gillies, Duncan F. 1994. Probabilistic Reasoning in High-Level Vision. *Image and Vision Computing*, **12**(1), 42–60. pages 285

Szeliski, Richard, Zabih, Ramin, Scharstein, Daniel, Veksler, Olga, Kolmogorov, Vladimir, Agarwala, Aseem, Tappen, Marshall, and Rother, Carsten. 2008. A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-based Priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**(6), 1068–1080. pages 285

Tandra, Haryono. 2014. The Relationship Between the Change of Variable Theorem and The Fundamental Theorem of Calculus for the Lebesgue Integral. *Teaching of Mathematics*, **17**(2), 76–83. pages 221

Tenenbaum, Joshua B., De Silva, Vin, and Langford, John C. 2000. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, **290**(5500), 2319–2323. pages 138

Tibshirani, Robert. 1996. Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society B*, **58**(1), 267–288. pages 305, 318

Tipping, Michael E., and Bishop, Christopher M. 1999. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B*, **61**(3), 611–622. pages 203, 342, 346, 347

Titsias, Michalis K., and Lawrence, Neil D. 2010. Bayesian Gaussian Process Latent Variable Model. In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*. JMLR W&CP, vol. 9. pages 349

Toussaint, Marc. 2012. *Some Notes on Gradient Descent*. pages 233, 250

Trefethen, Lloyd N., and Bau III, David. 1997. *Numerical Linear Algebra*. Vol. 50. SIAM. pages 139, 233

Tucker, Ledyard R. 1966. Some Mathematical Notes on Three-mode Factor Analysis. *Psychometrika*, **31**(3), 279–311. pages 139

Vapnik, Vladimir N. 1998. *Statistical Learning Theory*. Wiley. pages 268

©2018 Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong. To be published by Cambridge University Press.

Vapnik, Vladimir N. 1999. An Overview of Statistical Learning Theory. *IEEE Transactions on Neural Networks*, **10**(5), 988–999. pages 268

Vapnik, Vladimir N. 2000. *The Nature of Statistical Learning Theory*. Springer. pages 378, 397

Vishwanathan, S.V.N., Schraudolph, Nicol N., Kondor, Risi, and Borgwardt, Karsten M. 2010. Graph Kernels. *Journal of Machine Learning Research*, **11**, 1201–1242. pages 393

von Luxburg, Ulrike, and Schölkopf, Bernhard. 2011. Statistical Learning Theory: Models, Concepts, and Results. Pages 651–706 of: *Handbook of the History of Logic*, vol. 10. Elsevier. pages 268

Wahba, Grace. 1990. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics. pages 397

Walpole, Ronald E., Myers, Raymond H., Myers, Sharon L., and Ye, Keying. 2011. *Probability & Statistics for Engineers & Scientists*. 9th edn. Prentice Hall. pages 179, 225

Wasserman, Larry. 2004. *All of Statistics*. Springer. pages 219

Wasserman, Larry. 2007. *All of Nonparametric Statistics*. Springer. pages 214

Whittle, Peter. 2000. *Probability via Expectation*. Springer. pages 191, 225

Wickham, Hadley. 2014. Tidy Data. *Journal of Statistical Software*, **59**. pages 256

Williams, Christopher K. I. 1997. Computing with Infinite Networks. In: *Advances in Neural Information Processing Systems*. pages 318

Yu, Yaoliang, Cheng, Hao, Schuurmans, Dale, and Szepesvári, Csaba. 2013. Characterizing the Representer Theorem. In: *Proceedings of the International Conference on Machine Learning*. pages 387

Zhang, Haizhang, Xu, Yuesheng, and Zhang, Jun. 2009. Reproducing Kernel Banach Spaces for Machine Learning. *Journal of Machine Learning Research*, **10**, 2741–2775. pages 397

Zia, Royce K. P., Redish, Edward F., and McKay, Susan R. 2009. Making Sense of the Legendre Transform. *American Journal of Physics*, **77**(614). pages 251