
References

- 6640 Abel, Niels H. 1826. *Démonstration de l’Impossibilité de la Résolution Algébrique des*
6641 *Équations Générales qui Passent le Quatrième Degré*. Grøndahl & Søn. pages 315
- 6642 Adhikari, Ani, and DeNero, John. 2018. *Computational and Inferential Thinking: The*
6643 *Foundations of Data Science*. Gitbooks. pages 238
- 6644 Agarwal, Arvind, and III, Hal Daumé. 2010. A Geometric View of Conjugate Priors.
6645 *Machine Learning*, **81**(1), 99–113. pages 205
- 6646 Agresti, A. 2002. *Categorical Data Analysis*. Wiley. pages 254
- 6647 Akaike, Hirotugu. 1974. A New Look at the Statistical Model Identification. *IEEE*
6648 *Transactions on Automatic Control*, **19**(6), 716–723. pages 268
- 6649 Akhiezer, N.I., and Glazman, I.M. 1993. *Theory of Linear Operators in Hilbert Space*.
6650 Dover Publications, Inc. pages 371
- 6651 Alpaydin, Ethem. 2010. *Introduction to Machine Learning*. The MIT Press. pages 2
- 6652 Amari, Shun-ichi. 2016. *Information Geometry and Its Applications*. Springer. pages
6653 188
- 6654 Argyriou, Andreas, and Dinuzzo, Francesco. 2014. A Unifying View of Representer
6655 Theorems. Pages 748–756 of: Xing, Eric P., and Jebara, Tony (eds), *Proceedings*
6656 *of the 31st International Conference on Machine Learning*. Proceedings of Machine
6657 Learning Research, vol. 32, no. 2. Beijing, China: PMLR. pages 363
- 6658 Aronszajn, N. 1950. Theory of Reproducing Kernels. *Transactions of the American*
6659 *Mathematical Society*, **68**, 337–404. pages 367, 371
- 6660 Axler, Sheldon. 2015. *Linear Algebra Done Right*. third edn. Springer. pages 90
- 6661 Bakir, Gökhan, Hofmann, Thomas, Schölkopf, Bernhard, Smola, Alexander J., Taskar,
6662 Ben, and Vishwanathan, S.V.N (eds). 2007. *Predicting Structured Data*. MIT Press.
6663 pages 263
- 6664 Barber, David. 2012. *Bayesian Reasoning and Machine Learning*. Cambridge University
6665 Press. pages 2, 210, 258, 294, 347
- 6666 Barndorff-Nielsen, Ole. 2014. *Information and Exponential Families: In Statistical The-*
6667 *ory*. John Wiley and Sons. pages 209
- 6668 Bartholomew, David, Knott, Martin, and Moustaki, Irini. 2011. *Latent Variable Models*
6669 *and Factor Analysis: A Unified Approach*. Wiley & Sons. pages 327
- 6670 Beck, Amir, and Teboulle, Marc. 2003. Mirror descent and nonlinear projected sub-
6671 gradient methods for convex optimization. *Operations Research Letters*, **31**(3), 167–
6672 175. pages 233
- 6673 Belabbas, Mohamed-Ali, and Wolfe, Patrick J. 2009. Spectral methods in machine
6674 learning and new strategies for very large datasets. *Proceedings of the National*
6675 *Academy of Sciences*, pnas–0810600105. pages 133
- 6676 Belkin, Mikhail, and Niyogi, Partha. 2003. Laplacian eigenmaps for dimensionality
6677 reduction and data representation. *Neural computation*, **15**(6), 1373–1396. pages
6678 133

- 6679 Ben-Hur, Asa, Ong, Cheng Soon, Sonnenburg, Sören, Schölkopf, Bernhard, and Rätsch,
6680 Gunnar. 2008. Support Vector Machines and Kernels for Computational Biology.
6681 *PLoS Computational Biology*, 4(10), e1000173. pages 369, 371
- 6682 Bennett, Kristin P., and Bredensteiner, Erin J. 2000a. Duality and Geometry in SVM
6683 Classifiers. In: *Proceedings of the Seventeenth International Conference on Machine*
6684 *Learning*. pages 366
- 6685 Bennett, Kristin P., and Bredensteiner, Erin J. 2000b. Geometry in Learning. In: Gorini,
6686 Catherine A. (ed), *In Geometry at Work*. The Mathematical Association of America.
6687 pages 366
- 6688 Berlinet, Alain, and Thomas-Agnan, Christine. 2004. *Reproducing Kernel Hilbert Spaces*
6689 *in Probability and Statistics*. Springer. pages 367
- 6690 Bertsekas, Dimitri P. 1999. *Nonlinear Programming*. Athena Scientific. pages 233
- 6691 Bertsekas, Dimitri P. 2009. *Convex Optimization Theory*. Athena Scientific. pages 233
- 6692 Bickel, Peter J., and Doksum, Kjell. 2006. *Mathematical Statistics, Basic Ideas and*
6693 *Selected Topics, Vol I*. Prentice Hall. pages 210
- 6694 Bickson, Danny, Dolev, Danny, Shental, Ori, Siegel, Paul H., and Wolf, Jack K. 2007.
6695 Linear Detection via Belief Propagation. In: *Proceedings of the Annual Allerton Con-*
6696 *ference on Communication, Control, and Computing*. pages 263
- 6697 Billingsley, Patrick. 1995. *Probability and Measure*. Wiley. pages 176, 210
- 6698 Bishop, Christopher M. 1995. *Neural Networks for Pattern Recognition*. Clarendon
6699 Press. pages 295
- 6700 Bishop, Christopher M. 1999. Bayesian PCA. Pages 382–388 of: *Advances in Neural*
6701 *Information Processing Systems*. pages 326
- 6702 Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Information
6703 Science and Statistics. Springer-Verlag. pages vii, 2, 90, 167, 171, 173, 184, 206,
6704 210, 258, 259, 263, 279, 294, 346, 347, 371, 372
- 6705 Blei, David M., Kucukelbir, Alp, and McAuliffe, Jon D. 2017. Variational Inference: A
6706 Review for Statisticians. *Journal of the American Statistical Association*, 112(518),
6707 859–877. pages 258, 326
- 6708 Bonnans, J. Frédéric, Gilbert, J. Charles, Lemaréchal, Claude, and Sagastizábal, Clau-
6709 dia A. 2006. *Numerical Optimization: Theoretical and Practical Aspects*. 2nd edn.
6710 Springer Verlag. pages 233
- 6711 Borwein, Jonathan M., and Lewis, Adrian S. 2006. *Convex Analysis and Nonlinear*
6712 *Optimization*. 2nd edn. Canadian Mathematical Society. pages 233
- 6713 Bottou, Leon. 1998. Online Algorithms and Stochastic Approximations. Pages 1–34
6714 of: *Online Learning and Neural Networks*. Cambridge University Press. pages 219
- 6715 Bottou, Léon, Curtis, Frank E., and Nocedal, Jorge. 2018. Optimization Methods for
6716 Large-Scale Machine Learning. *ArXiv*. pages 219, 233
- 6717 Boucheron, Stephane, Lugosi, Gabor, and Massart, Pascal. 2013. *Concentration In-*
6718 *equalities: A Nonasymptotic Theory of Independence*. Oxford University Press. pages
6719 173
- 6720 Boyd, Stephen, and Vandenberghe, Lieven. 2004. *Convex Optimization*. Cambridge
6721 University Press. pages 217, 221, 223, 233
- 6722 Boyd, Stephen, and Vandenberghe, Lieven. 2018. *Introduction to Applied Linear Alge-*
6723 *bra*. Cambridge University Press. pages 90
- 6724 Brooks, Steve, Gelman, Andrew, Jones, Galin L., and Meng, Xiao-Li (eds). 2011. *Hand-*
6725 *book of Markov Chain Monte Carlo*. Chapman and Hall/CRC. pages 258
- 6726 Brown, Lawrence D. 1986. *Fundamentals of Statistical Exponential Families: with Ap-*
6727 *plications in Statistical Decision Theory*. Lecture Notes - Monograph Series. Institute
6728 of Mathematical Statistics. pages 207
- 6729 Bryson, Arthur E. 1961. A Gradient Method for Optimizing Multi-stage Allocation
6730 Processes. In: *Proceedings of the Harvard University Symposium on Digital Computers*
6731 *and Their Applications*. pages 156, 158

- 6732 Bubeck, Sébastien. 2015. Convex Optimization: Algorithms and Complexity. *Foundations and Trends® in Machine Learning*, **8**(3-4), 231–357. pages 233
- 6733
- 6734 Bühlmann, Peter, and Geer, Sara Van De. 2011. *Statistics for High-Dimensional Data*. Springer. pages 250
- 6735
- 6736 Burges, Christopher. 2010. Dimension Reduction: A Guided Tour. *Foundations and Trends in Machine Learning*, **2**(4), 275–365. pages 327
- 6737
- 6738 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 133, 134
- 6739
- 6740 Casella, George, and Berger, Roger L. 2002. *Statistical Inference*. Duxbury. pages 182, 188, 191, 255
- 6741
- 6742
- 6743 Çınlar, Erhan. 2011. *Probability and Stochastics*. Springer. pages 210
- 6744
- 6745 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 370
- 6746
- 6747 Cheeseman, Peter. 1985. In Defense of Probability. In: *IJCAI*. pages 255
- 6748
- 6749 Chollet, Francois, and Allaire, J. J. 2018. *Deep Learning with R*. Manning Publications. pages 2
- 6750
- 6751 Codd, Edgar F. 1990. *The Relational Model for Database Management*. Addison-Wesley Longman Publishing. pages 238
- 6752
- 6753 Cunningham, John P., and Ghahramani, Zoubin. 2015. Linear Dimensionality Reduction: Survey, Insights, and Generalizations. *Journal of Machine Learning Research*, **16**, 2859–2900. pages 327
- 6754
- 6755 Datta, Biswa Nath. 2010. *Numerical Linear Algebra and Applications*. Vol. 116. Siam. pages 125
- 6756
- 6757 Davidson, Anthony C., and Hinkley, David V. 1997. *Bootstrap Methods and their Application*. Cambridge University Press. pages 250
- 6758
- 6759 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. Pages 1–11 of: *Advances in Neural Information Processing Systems*. pages 219
- 6760
- 6761
- 6762
- 6763 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 263
- 6764
- 6765
- 6766 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 196
- 6767
- 6768
- 6769 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 87
- 6770
- 6771
- 6772 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 341
- 6773
- 6774
- 6775 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 79
- 6776
- 6777
- 6778 Devroye, Luc. 1986. *Non-Uniform Random Variate Generation*. Springer-Verlag. pages 201
- 6779
- 6780 Domingos, Pedro. 2012. A few useful things to know about machine learning. *Communications of the ACM*, **55**, 78–87. pages 12
- 6781
- 6782 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 133
- 6783
- 6784

- 6785 Dostál, Zdeněk. 2009. *Optimal Quadratic Programming Algorithms: With Applications*
6786 *to Variational Inequalities*. Springer-Verlag. pages 364
- 6787 Douven, Igor. 2017. Abduction. In: Zalta, Edward N. (ed), *The Stanford Encyclopedia*
6788 *of Philosophy*, summer 2017 edn. Metaphysics Research Lab, Stanford University.
6789 pages 243
- 6790 Downey, Allen B. 2014. *Think Stats: Exploratory Data Analysis*. 2nd edn. O'Reilly
6791 Media. pages 209
- 6792 Dreyfus, Stuart. 1962. The Numerical Solution of Variational Problems. *Journal of*
6793 *Mathematical Analysis and Applications*, 5(1), 30–45. pages 156, 158
- 6794 Drumm, Volker, and Weil, Wolfgang. 2001. *Lineare Algebra und Analytische Geometrie*.
6795 Lecture Notes, Universität Karlsruhe. pages 18, 53
- 6796 Dudley, R. M. 2002. *Real Analysis and Probability*. Cambridge University Press. pages
6797 210
- 6798 Efron, Bradley, and Hastie, Trevor. 2016. *Computer Age Statistical Inference: Algorithms,*
6799 *Evidence and Data Science*. Cambridge University Press. pages 202, 253, 255
- 6800 Efron, Bradley, and Tibshirani, Robert J. 1993. *An Introduction to the Bootstrap*. Chap-
6801 man and Hall/CRC. pages 250
- 6802 Elliott, Conal. 2009. Beautiful differentiation. In: *International Conference on Func-*
6803 *tional Programming (ICFP)*. pages 166
- 6804 Evgeniou, Theodoros, Pontil, Massimiliano, and Poggio, Tomaso. 2000. Statistical
6805 Learning Theory: A Primer. *International Journal of Computer Vision*, 38(1), 9–13.
6806 pages 249
- 6807 Fan, Rong-En, Chang, Kai-Wei, Hsieh, Cho-Jui, Wang, Xiang-Rui, and Lin, Chih-Jen.
6808 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learn-*
6809 *ing Research*, 9, 1871–1874. pages 359
- 6810 Gal, Yarín, van der Wilk, Mark, and Rasmussen, Carl E. 2014. Distributed Variational
6811 Inference in Sparse Gaussian Process Regression and Latent Variable Models. In:
6812 *Advances in Neural Information Processing Systems*. pages 219
- 6813 Gärtner, Thomas. 2008. *Kernels for Structured Data*. World Scientific. pages 369
- 6814 Gavish, Matan, and Donoho, David L. 2014. The Optimal Hard Threshold for Singular
6815 Values is $4\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8), 5040–5053. pages
6816 326
- 6817 Ghahramani, Zoubin. 2015. Probabilistic Machine Learning and Artificial Intelligence.
6818 *Nature*, 521, 452–459. pages 178, 258
- 6819 Ghahramani, Zoubin, and Roweis, Sam T. 1999. Learning Nonlinear Dynamical Sys-
6820 tems using an EM Algorithm. In: Kearns, M. S., Solla, S. A., and Cohn, D. A. (eds),
6821 *Advances in Neural Information Processing Systems*, vol. 11. The MIT Press. pages
6822 347
- 6823 Gilks, Walter R., Richardson, Sylvia, and Spiegelhalter, David J. 1996. *Markov Chain*
6824 *Monte Carlo in Practice*. Chapman & Hall. pages 326
- 6825 Gneiting, Tilmann, and Raftery, Adrian E. 2007. Strictly proper scoring rules, pre-
6826 diction, and estimation. *Journal of the American Statistical Association*, 102(477),
6827 359–378. pages 372
- 6828 Goh, Gabriel. 2017. Why Momentum Really Works. *Distill*. pages 217, 233
- 6829 Gohberg, Israel, Goldberg, Seymour, and Krupnik, Nahum. 2012. *Traces and determi-*
6830 *nants of linear operators*. Vol. 116. Birkhäuser. pages 99
- 6831 Gonçalves, Hugo. 2014. *Legendre and Legendre-Fenchel transforms*. Accessed on 3
6832 March 2018. pages 233
- 6833 Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron. 2016. *Deep Learning*. MIT
6834 Press. <http://www.deeplearningbook.org>. pages 210, 240, 263, 295
- 6835 Griewank, Andreas, and Walther, Andrea. 2003. Introduction to Automatic Differenti-
6836 ation. *PAMM*, 2(1), 45–49. pages 166

- 6837 Griewank, Andreas, and Walther, Andrea. 2008. *Evaluating Derivatives, Principles and*
6838 *Techniques of Algorithmic Differentiation*. second edn. SIAM, Philadelphia. pages
6839 166
- 6840 Grinstead, Charles M., and Snell, J. Laurie. 1997. *Introduction to Probability*. American
6841 Mathematical Society. pages 172, 194, 209
- 6842 Hacking, Ian. 2001. *Probability and Inductive Logic*. Cambridge University Press. pages
6843 209
- 6844 Hall, Peter. 1992. *The Bootstrap and Edgeworth Expansion*. Springer. pages 250
- 6845 Hasselblatt, Boris, and Katok, Anatole. 2003. *A first course in dynamics with a*
6846 *Panorama of Recent Developments*. Cambridge University Press. pages 172
- 6847 Hazan, Elad. 2015. Introduction to Online Convex Optimization. *Foundations and*
6848 *Trends*, **2**(3-4), 157–325. pages 233
- 6849 Hensman, James, Fusi, Nicolò, and Lawrence, Neil D. 2013. Gaussian Processes for
6850 Big Data. In: Nicholson, A., and Smyth, P. (eds), *Proceedings of the Conference on*
6851 *Uncertainty in Artificial Intelligence*. AUAI Press. pages 219
- 6852 Herbrich, Ralf, Minka, Tom, and Graepel, Thore. 2007. TrueSkill(TM): A Bayesian
6853 Skill Rating System. Pages 569–576 of: *Advances in Neural Information Processing*
6854 *Systems*. MIT Press. pages 263
- 6855 Hiriart-Urruty, Jean-Baptiste, and Lemaréchal, Claude. 2001. *Fundamentals of Convex*
6856 *Analysis*. Springer. pages 229, 233
- 6857 Hoffman, Matthew D., Blei, David M., Wang, Chong, and Paisley, John. 2013. Stochas-
6858 tic Variational Inference. *Journal of Machine Learning Research*, **14**(1), 1303–1347.
6859 pages 219
- 6860 Hofmann, Thomas, Schölkopf, Bernhard, and Smola, Alexander J. 2008. Kernel meth-
6861 ods in machine learning. *Ann. Statist.*, **36**(3), 1171–1220. pages 363
- 6862 Hogben, L. (ed). 2013. *Handbook of Linear Algebra*. 2nd edn. Discrete Mathematics
6863 and Its Applications. Chapman and Hall. pages 18
- 6864 Hogben, Leslie. 2006. *Handbook of linear algebra*. Chapman and Hall/CRC. pages 103
- 6865 Hotelling, Harold. 1933. Analysis of a Complex of Statistical Variables into Principal
6866 Components. *Journal of Educational Psychology*, **24**, 417–441. pages 79, 90, 297,
6867 301
- 6868 Imbens, Guido W., and Rubin, Donald B. 2015. *Causal Inference for statistics, social*
6869 *and biomedical sciences*. Cambridge University Press. pages 263
- 6870 Jacod, Jean, and Protter, Philip. 2004. *Probability Essentials*. 2nd edn. Springer-Verlag.
6871 pages 171, 210
- 6872 Jaynes, Edwin T. 2003. *Probability Theory: The Logic of Science*. Cambridge University
6873 Press. pages 169, 170, 172, 178, 179, 210, 255
- 6874 Jefferys, Willian H., and Berger, James O. 1992. Ockham’s Razor and Bayesian Analy-
6875 sis. *American Scientist*, **80**, 64–72. pages 265
- 6876 Joachims, Thorsten. 1999. Making large-Scale SVM Learning Practical. In: Schölkopf,
6877 B., Burges, C., and Smola, A. (eds), *Advances in Kernel Methods - Support Vector*
6878 *Learning*. MIT Press. pages 370
- 6879 Julier, Simon J., and Uhlmann, Jeffrey K. 1997. A New Extension of the Kalman Filter
6880 to Nonlinear Systems. Pages 182–193 of: *Proceedings of AeroSense: 11th Symposium*
6881 *on Aerospace/Defense Sensing, Simulation and Controls*. pages 167
- 6882 Kalman, Dan. 1996. A singularly valuable decomposition: the SVD of a matrix. *The*
6883 *College Mathematics Journal*, **27**(1), 2–23. pages 115
- 6884 Kalman, Rudolf E. 1960. A New Approach to Linear Filtering and Prediction Problems.
6885 *Transactions of the ASME—Journal of Basic Engineering*, **82**(Series D), 35–45. pages
6886 196
- 6887 Katz, Victor J. 2004. *A History of Mathematics*. Pearson/Addison-Wesley. pages 105
- 6888 Kelley, Henry J. 1960. Gradient Theory of Optimal Flight Paths. *Ars Journal*, **30**(10),
6889 947–954. pages 156, 158

- 6890 Kimeldorf, George S., and Wahba, Grace. 1970. A correspondence between Bayesian
6891 estimation on stochastic processes and smoothing by splines. *The Annals of Mathe-*
6892 *matical Statistics*, **41**(2), 495–502. pages 363
- 6893 Kittler, J., and Föglein, J. 1984. Contextual Classification of Multispectral Pixel Data.
6894 *Image and Vision Computing*, **2**(1), 13–29. pages 263
- 6895 Kolda, Tamara G, and Bader, Brett W. 2009. Tensor decompositions and applications.
6896 *SIAM review*, **51**(3), 455–500. pages 134
- 6897 Koller, Daphne, and Friedman, Nir. 2009. *Probabilistic Graphical Models*. MIT Press.
6898 pages 263
- 6899 Lang, Serge. 1987. *Linear Algebra*. Springer-Verlag, New York. pages 111
- 6900 Lawrence, Neil. 2005. Probabilistic Non-linear Principal Component Analysis with
6901 Gaussian Process Latent Variable Models. *Journal of Machine Learning Research*,
6902 **6**(Nov.), 1783–1816. pages 328
- 6903 Leemis, Lawrence M., and McQueston, Jacquelyn T. 2008. Univariate Distribution
6904 Relationships. *The American Statistician*, **62**(1), 45–53. pages 202, 204
- 6905 Lehmann, Erich L., and Romano, Joseph P. 2005. *Testing Statistical Hypotheses*.
6906 Springer. pages 191
- 6907 Lehmann, Erich Leo, and Casella, George. 1998. *Theory of Point Estimation*. Springer.
6908 pages 207, 210, 253
- 6909 Liesen, Jörg, and Mehrmann, Volker. 2015. *Linear Algebra*. Springer Undergraduate
6910 Mathematics Series. Springer. pages 18
- 6911 Loosli, Gaëlle, Canu, Stéphane, and Ong, Cheng Soon. 2016. Learning SVM in Kreĭn
6912 Spaces. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, **38**(6), 1204–
6913 1216. pages 371
- 6914 Luenberger, David G. 1969. *Optimization by Vector Space Methods*. John Wiley and
6915 Sons. pages 233
- 6916 MacKay, David J. C. 2003a. *Information Theory, Inference, and Learning Algorithms*. The
6917 Edinburgh Building, Cambridge CB2 2RU, UK: Cambridge University Press. pages
6918 265, 325, 347
- 6919 MacKay, David J. C. 1992. Bayesian Interpolation. *Neural Computation*, **4**, 415–447.
6920 pages 265
- 6921 MacKay, David J. C. 1998. Introduction to Gaussian Processes. Pages 133–165 of:
6922 Bishop, C. M. (ed), *Neural Networks and Machine Learning*, vol. 168. Berlin, Ger-
6923 many: Springer. pages 296
- 6924 MacKay, David J. C. 2003b. *Information Theory, Inference and Learning Algorithms*.
6925 Cambridge University Press. pages 2, 210
- 6926 Magnus, Jan R., and Neudecker, Heinz. 2007. *Matrix Differential Calculus with Appli-*
6927 *cations in Statistics and Econometrics*. 3rd edn. John Wiley & Sons. pages 166
- 6928 Manton, Jonathan H., and Amblard, Pierre-Olivier. 2015. A Primer on Reproducing
6929 Kernel Hilbert Spaces. *Foundations and Trends in Signal Processing*, **8**(1–2), 1–126.
6930 pages 371
- 6931 Markovsky, Ivan. 2011. *Low rank approximation: algorithms, implementation, applica-*
6932 *tions*. Springer Science & Business Media. pages 134
- 6933 Maybeck, Peter S. 1979. *Stochastic Models, Estimation, and Control*. Mathematics in
6934 Science and Engineering, vol. 141. Academic Press, Inc. pages 167
- 6935 McCullagh, Peter, and Nelder, John A. 1989. *Generalized Linear Models*. second edn.
6936 CRC Press. pages 254
- 6937 McEliece, Robert J., MacKay, David J. C., and Cheng, Jung-Fu. 1998. Turbo Decoding
6938 as an Instance of Pearl’s “Belief Propagation” Algorithm. *IEEE Journal on Selected*
6939 *Areas in Communications*, **16**(2), 140–152. pages 263
- 6940 Meyer, Carl D. 2000. *Matrix Analysis and Applied Linear Algebra*. Vol. 71. SIAM. pages
6941 103

- 6942 Mika, Sebastian, Rätsch, Gunnar, Weston, Jason, Schölkopf, Bernhard, and Müller,
6943 Klaus-Robert. 1999. Fisher discriminant analysis with kernels. Pages 41–48 of:
6944 *Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal*
6945 *processing society workshop*. Ieee. pages 133
- 6946 Minka, Tom. 2001. Automatic Choice of dimensionality of PCA. In: *Neural Information*
6947 *Processing Systems (NIPS)*. pages 326
- 6948 Mitchell, Tom. 1997. *Machine Learning*. McGraw Hill. pages 247
- 6949 Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Rusu, Andrei A., Veness, Joel,
6950 Bellemare, Marc G., Graves, Alex, Riedmiller, Martin, Fidjeland, Andreas K., Os-
6951 trovski, Georg, Petersen, Stig, Beattie, Charles, Sadik, Amir, Antonoglou, Ioan-
6952 nis, King, Helen, Kumaran, Dhharshan, Wierstra, Daan, Legg, Shane, and Hassabis,
6953 Demis. 2015. Human-Level Control through Deep Reinforcement Learning. *Nature*,
6954 **518**(Feb.), 529–533. pages 219
- 6955 Moonen, Marc, and De Moor, Bart. 1995. *SVD and Signal Processing, III: Algorithms,*
6956 *Architectures and Applications*. Elsevier. pages 134
- 6957 Müller, Andreas C., and Guido, Sarah. 2016. *Introduction to Machine Learning with*
6958 *Python: A Guide for Data Scientists*. O'Reilly Publishing. pages 2
- 6959 Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA,
6960 USA: MIT Press. pages 2, 197, 204, 210, 258, 265, 267, 268, 294, 327, 347, 371
- 6961 Neal, Radford M. 1996. *Bayesian Learning for Neural Networks*. Ph.D. thesis, Depart-
6962 ment of Computer Science, University of Toronto. pages 296
- 6963 Neal, Radford M., and Hinton, Geoffrey E. 1999. A View of the EM Algorithm that
6964 Justifies Incremental, Sparse, and Other Variants. Pages 355–368 of: Jordan, M. I.
6965 (ed), *Learning in Graphical Models*. MIT Press. pages 341
- 6966 Nelsen, Roger. 2006. *An Introduction to Copulas*. Springer. pages 191
- 6967 Neumaier, Arnold. 1998. Solving Ill-Conditioned and Singular Linear Systems: A Tu-
6968 torial on Regularization. *SIAM Review*, **40**, 636–666. pages 250
- 6969 Nocedal, Jorge, and Wright, Stephen J. 2006. *Numerical Optimization*. Springer Series
6970 in Operations Research. Springer. pages 161, 233, 371
- 6971 Nowozin, Sebastian, Gehler, Peter V., Jancsary, Jeremy, and Lampert, Christoph H.
6972 (eds). 2014. *Advanced Structured Prediction*. MIT Press. pages 263
- 6973 O'Hagan, Anthony. 1991. Bayes-Hermite Quadrature. *Journal of Statistical Planning*
6974 *and Inference*, **29**, 245–260. pages 268
- 6975 Ong, Cheng Soon, Mary, Xavier, Canu, Stéphane, , and Smola, Alexander J. 2004.
6976 Learning with non-positive kernels. Pages 639–646 of: *International Conference on*
6977 *Machine Learning (ICML)*. pages 371
- 6978 Ormonoit, Dirk, Sidenbladh, Hedvig, Black, Michael J, and Hastie, Trevor. 2001. Learn-
6979 ing and tracking cyclic human motion. Pages 894–900 of: *Advances in Neural Infor-*
6980 *mation Processing Systems*. pages 134
- 6981 Page, Lawrence, Brin, Sergey, Motwani, Rajeev, and Winograd, Terry. 1999. *The PageR-*
6982 *ank Citation Ranking: Bringing Order to the Web*. Tech. rept. Stanford InfoLab. pages
6983 108, 315
- 6984 Parzen, Emanuel. 1962. On Estimation of a Probability Density Function and Mode.
6985 *The Annals of Mathematical Statistics*, **33**(3), 1065–1076. pages 348
- 6986 Pearl, Judea. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible*
6987 *Inference*. Morgan Kaufmann. pages 170, 262
- 6988 Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference*. 2nd edn. Cambridge
6989 University Press. pages 259, 263
- 6990 Pearson, Karl. 1895. Contributions to the Mathematical Theory of Evolution. II. Skew
6991 Variation in Homogeneous Material. *Philosophical Transactions of the Royal Society*
6992 *A: Mathematical, Physical and Engineering Sciences*, **186**, 343–414. pages 348

- 6993 Pearson, Karl. 1901a. LIII. On lines and planes of closest fit to systems of points in
6994 space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of*
6995 *Science*, 2(11), 559–572. pages 133
- 6996 Pearson, Karl. 1901b. On Lines and Planes of Closest Fit to Systems of Points in Space.
6997 *Philosophical Magazine*, 2(11), 559–572. pages 78, 90, 297, 307
- 6998 Peters, Jonas, Janzing, Dominik, and Schölkopf, Bernhard. 2017. *Elements of Causal*
6999 *Inference: Foundations and Learning Algorithms*. The MIT Press. pages 263
- 7000 Petersen, K. B., and Pedersen, M. S. 2012 (nov). *The Matrix Cookbook*. Tech. rept.
7001 Technical University of Denmark. Version 20121115. pages 155
- 7002 Pollard, David. 2002. *A User’s Guide to Measure Theoretic Probability*. Cambridge
7003 University Press. pages 210
- 7004 Polyak, Roman A. 2016. The Legendre Transformation in Modern Optimization. Pages
7005 437–507 of: *Optimization and Its Applications in Control and Data Sciences*. Springer
7006 *Optimization and Its Applications*. pages 233
- 7007 Press, William H., Teukolsky, Saul A., Vetterling, William T., and Flannery, Brian P.
7008 2007. *Numerical Recipes: The Art of Scientific Computing*. third edn. Cambridge
7009 University Press. pages 125, 133
- 7010 Raschka, Sebastian, and Mirjalili, Vahid. 2017. *Python Machine Learning: Machine*
7011 *Learning and Deep Learning with Python, scikit-learn, and TensorFlow*. Packt Publish-
7012 ing. pages 2
- 7013 Rasmussen, Carl E., and Ghahramani, Zoubin. 2003. Bayesian Monte Carlo. Pages
7014 489–496 of: Becker, S., Thrun, S., and Obermayer, K. (eds), *Advances in Neural*
7015 *Information Processing Systems 15*. Cambridge, MA, USA: The MIT Press. pages 268
- 7016 Rasmussen, Carl E., and Williams, Christopher K. I. 2006. *Gaussian Processes for Ma-*
7017 *chine Learning*. Adaptive Computation and Machine Learning. Cambridge, MA,
7018 USA: The MIT Press. pages 90, 196, 210, 295, 296
- 7019 Rasmussen, Carl Edward, and Ghahramani, Zoubin. 2001. Occam’s Razor. Pages 294–
7020 300 of: *Advances in Neural Information Processing Systems 13*. The MIT Press. pages
7021 268
- 7022 Reid, Mark, and Williamson, Robert C. 2011. Information, Divergence and Risk for
7023 Binary Experiments. *Journal of Machine Learning Research*, 12, 731–817. pages 372
- 7024 Rezende, Danilo Jimenez, and Mohamed, Shakir. 2015. Variational Inference with
7025 Normalizing Flows. In: *International Conference on Machine Learning*. pages 209
- 7026 Rifkin, Ryan M., and Lippert, Ross A. 2007. Value Regularization and Fenchel Duality.
7027 *Journal of Machine Learning Research*, 8, 441–479. pages 371
- 7028 Rockafellar, R. Tyrrell. 1970. *Convex Analysis*. Princeton University Press. pages 233
- 7029 Rogers, Simon, and Girolami, Mark. 2016. *A First Course in Machine Learning*. Second
7030 edn. Chapman and Hall/CRC. pages 2, 346
- 7031 Rosenbaum, Paul R. 2017. *Observation & Experiment: An Introduction to Causal In-*
7032 *ference*. Harvard University Press. pages 263
- 7033 Rosenblatt, Murray. 1956. Remarks on Some Nonparametric Estimates of a Density
7034 Function. *The Annals of Mathematical Statistics*, 27(3), 832–837. pages 348
- 7035 Roweis, Sam, and Ghahramani, Zoubin. 1999. A Unifying Review of Linear Gaussian
7036 Models. *Neural Computation*, 11(2), 305–345. pages 197, 347
- 7037 Roweis, Sam T. 1998. EM Algorithms for PCA and SPCA. Pages 626–632 of: *Advances*
7038 *in Neural Information Processing Systems*. pages 326
- 7039 Roy, Anindya, and Banerjee, Sudipto. 2014. *Linear algebra and matrix analysis for*
7040 *statistics*. Chapman and Hall/CRC. pages 115
- 7041 Rubinstein, Reuven Y, and Kroese, Dirk P. 2016. *Simulation and the Monte Carlo*
7042 *method*. Vol. 10. John Wiley & Sons. pages 133
- 7043 Ruffini, Paolo. 1799. *Teoria Generale delle Equazioni, in cui si Dimostra Impossibile la*
7044 *Soluzione Algebraica delle Equazioni Generali di Grado Superiore al Quarto*. Stampe-
7045 ria di S. Tommaso d’Aquino. pages 315

- 7046 Rumelhart, David E., Hinton, Geoffrey E., and Williams, Ronald J. 1986. Learning
7047 Representations by Back-propagating Errors. *Nature*, **323**(6088), 533–536. pages
7048 156, 158, 217
- 7049 Saitoh, Saburo. 1988. *Theory of Reproducing Kernels and its Applications*. Longman
7050 Scientific & Technical. pages 367, 371
- 7051 Schölkopf, Bernhard, and Smola, Alexander J. 2002. *Learning with Kernels—Support
7052 Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation
7053 and Machine Learning. Cambridge, MA, USA: The MIT Press. pages 90, 295, 327,
7054 354, 364, 367, 371
- 7055 Schölkopf, Bernhard, Smola, Alexander, and Müller, Klaus-Robert. 1997. Kernel prin-
7056 cipal component analysis. Pages 583–588 of: *International Conference on Artificial
7057 Neural Networks*. Springer. pages 90
- 7058 Schölkopf, Bernhard, Smola, Alexander J., and Müller, Klaus-Robert. 1998. Nonlinear
7059 Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, **10**(5),
7060 1299–1319. pages 327
- 7061 Schölkopf, Bernhard, Herbrich, Ralf, and Smola, Alexander J. 2001. A generalized
7062 representer theorem. Pages 416–426 of: *International Conference on Computational
7063 Learning Theory (COLT)*. pages 363
- 7064 Schwartz, Laurent. 1964. Sous espaces Hilbertiens d’espaces vectoriels topologiques
7065 et noyaux associés. *Journal d’Analyse Mathématique*, **13**, 115–256. in French. pages
7066 371
- 7067 Schwarz, Gideon E. 1978. Estimating the Dimension of a Model. *Annals of Statistics*,
7068 **6**(2), 461–464. pages 268
- 7069 Shalev-Shwartz, Shai, and Ben-David, Shai. 2014. *Understanding Machine Learning:
7070 From Theory to Algorithms*. Cambridge University Press. pages 2, 173, 249, 355
- 7071 Shawe-Taylor, John, and Cristianini, Nello. 2004. *Kernel Methods for Pattern Analysis*.
7072 Cambridge University Press. pages 367, 371
- 7073 Shawe-Taylor, John, and Sun, Shiliang. 2011. A review of optimization methodologies
7074 in support vector machines. *Neurocomputing*, **74**(17), 3609–3618. pages 371
- 7075 Shental, O., Bickson, D., P. H. Siegel and, J. K. Wolf, and Dolev, D. 2008. Gaussian
7076 Belief Propagation Solver for Systems of Linear Equations. In: *IEEE International
7077 Symposium on Information Theory*. pages 263
- 7078 Shewchuk, Jonathan Richard. 1994 (August). *An Introduction to the Conjugate Gradi-
7079 ent Method Without the Agonizing Pain*. Carnegie Mellon University, Edition 1 1/4.
7080 pages 232
- 7081 Shi, Jianbo, and Malik, Jitendra. 2000. Normalized cuts and image segmentation. *IEEE
7082 Transactions on pattern analysis and machine intelligence*, **22**(8), 888–905. pages
7083 133
- 7084 Shi, Qinfeng, Petterson, James, Dror, Gideon, Langford, John, Smola, Alex, and Vish-
7085 wanathan, S.V.N. 2009. Hash Kernels for Structured Data. *Journal of Machine
7086 Learning Research*, 2615–2637. pages 369
- 7087 Shiriyayev, A. N. 1984. *Probability*. Springer. pages 210
- 7088 Shor, Naum Z. 1985. *Minimization Methods for Non-differentiable Functions*. Springer.
7089 pages 233
- 7090 Shotton, Jamie, Winn, John, Rother, Carsten, and Criminisi, Antonio. 2006. Texton-
7091 Boost: Joint Appearance, Shape and Context Modeling for Mult-Class Object Recog-
7092 nition and Segmentation. In: *Proceedings of the European Conference on Computer
7093 Vision*. pages 263
- 7094 Spearman, Charles. 1904. “General Intelligence,” Objectively Determined and Mea-
7095 sured. *The American Journal of Psychology*, **15**(2), 201–292. pages 327
- 7096 Spiegelhalter, David, and Smith, A. F. M. 1980. Bayes Factors and Choice Criteria for
7097 Linear Models. *Journal of the Royal Statistical Society B*, **42**(2), 213–220. pages 265

- 7098 Steinwart, Ingo. 2007. How to Compare Different Loss Functions and Their Risks.
7099 *Constructive Approximation*, **26**, 225–287. pages 371
- 7100 Steinwart, Ingo, and Christmann, Andreas. 2008. *Support Vector Machines*. Springer.
7101 pages 349, 353, 355, 359, 371
- 7102 Stoer, Josef, and Burlirsch, Roland. 2002. *Introduction to Numerical Analysis*. Springer.
7103 pages 90, 268
- 7104 Strang, Gilbert. 1993. The fundamental theorem of linear algebra. *The American
7105 Mathematical Monthly*, **100**(9), 848–855. pages 115
- 7106 Strang, Gilbert. 2003. *Introduction to Linear Algebra*. 3rd edn. Wellesley-Cambridge
7107 Press. pages 18, 76, 294
- 7108 Stray, Jonathan. 2016. *The Curious Journalist’s Guide to Data*. Tow Center for Digital
7109 Journalism at Columbia’s Graduate School of Journalism. pages 238
- 7110 Strogatz, Steven. 2014. Writing about Math for the Perplexed and the Traumatized.
7111 *Notices of the American Mathematical Society*, **61**(3), 286–291. pages 2
- 7112 Sucar, Luis E., and Gillies, Duncan F. 1994. Probabilistic Reasoning in High-Level
7113 Vision. *Image and Vision Computing*, **12**(1), 42–60. pages 263
- 7114 Szeliski, Richard, Zabih, Ramin, Scharstein, Daniel, Veksler, Olga, Kolmogorov,
7115 Vladimir, Agarwala, Aseem, Tappen, Marshall, and Rother, Carsten. 2008. A Com-
7116 parative Study of Energy Minimization Methods for Markov Random Fields with
7117 Smoothness-based Priors. *IEEE Transactions on Pattern Analysis and Machine Intelli-
7118 gence*, **30**(6), 1068–1080. pages 263
- 7119 Tenenbaum, Joshua B, De Silva, Vin, and Langford, John C. 2000. A global geometric
7120 framework for nonlinear dimensionality reduction. *science*, **290**(5500), 2319–2323.
7121 pages 133
- 7122 Tibshirani, Robert. 1996. Regression Selection and Shrinkage via the Lasso. *Journal
7123 of the Royal Statistical Society B*, **58**(1), 267–288. pages 280, 296
- 7124 Tipping, Michael E., and Bishop, Christopher M. 1999. Probabilistic Principal Com-
7125 ponent Analysis. *Journal of the Royal Statistical Society: Series B*, **61**(3), 611–622.
7126 pages 197, 320, 325
- 7127 Titsias, Michalis K., and Lawrence, Neil D. 2010. Bayesian Gaussian Process Latent
7128 Variable Model. Pages 844–851 of: Teh, Y. W., and Titterington, D. M. (eds), *Pro-
7129 ceedings of the International Conference on Artificial Intelligence and Statistics*. JMLR
7130 W&CP, vol. 9. pages 328
- 7131 Toussaint, Marc. 2012. *Some Notes on Gradient Descent*. pages 216, 233
- 7132 Trefethen, Lloyd N, and Bau III, David. 1997. *Numerical Linear Algebra*. Vol. 50. Siam.
7133 pages 134, 216
- 7134 Tucker, Ledyard R. 1966. Some mathematical notes on three-mode factor analysis.
7135 *Psychometrika*, **31**(3), 279–311. pages 134
- 7136 Vapnik, Vladimir. 2000. *The Nature of Statistical Learning Theory*. Springer Verlag.
7137 pages 355, 371
- 7138 Vapnik, Vladimir N. 1998. *Statistical Learning Theory*. Wiley. pages 249
- 7139 Vapnik, Vladimir N. 1999. An Overview of Statistical Learning Theory. *IEEE Transac-
7140 tions on Neural Networks*, **10**(5), 988–999. pages 249
- 7141 Vishwanathan, S.V.N., Schraudolph, Nicol N., Kondor, Risi, and Borgwardt, Karsten M.
7142 2010. Graph Kernels. *Journal of Machine Learning Research*, **11**, 1201–1242. pages
7143 369
- 7144 von Luxburg, Ulrike, and Schölkopf, Bernhard. 2011. *Statistical Learning Theory: Mod-
7145 els, Concepts, and Results*. Vol. 10. Amsterdam, Netherlands: Elsevier North Holland.
7146 Pages 651–706. pages 249
- 7147 Wahba, Grace. 1990. *Spline Models for Observational Data*. Society for Industrial and
7148 Applied Mathematics. pages 371
- 7149 Wasserman, Larry. 2004. *All of Statistics*. Springer. pages 189
- 7150 Wasserman, Larry. 2007. *All of Nonparametric Statistics*. Springer. pages 207

- 7151 Wickham, Hadley. 2014. Tidy Data. *Journal of Statistical Software*, **59**. pages 238
- 7152 Williams, Christopher K. I. 1997. Computing with Infinite Networks. In: *Neural Infor-*
7153 *mation Processing Systems*. pages 296
- 7154 Yu, Yaoliang, Cheng, Hao, Schuurmans, Dale, and Szepesvári, Csaba. 2013. Charac-
7155 terizing the representer theorem. Pages 570–578 of: *International Conference on*
7156 *Machine Learning (ICML)*. pages 363
- 7157 Zhang, Haizhang, Xu, Yuesheng, and Zhang, Jun. 2009. Reproducing Kernel Banach
7158 Spaces for Machine Learning. *Journal of Machine Learning Research*, **10**, 2741–2775.
7159 pages 371
- 7160 Zia, Royce K. P., Redish, Edward F., and McKay, Susan R. 2009. Making sense of the
7161 Legendre transform. *American Journal of Physics*, **77**(614). pages 233