Risk bounds for deep learning

Bos, J.M.

Bibliography

[1] Aas, K., Czado, C., Frigessi, A., and Bakken, H. Pair-copula constructions of multiple dependence. *Insurance Math. Econom. 44*, 2 (2009), 182–198.

[2] Ahle, T. D. Sharp and simple bounds for the raw moments of the binomial and Poisson distributions. *Statist. Probab. Lett. 182* (2022), Paper No. 109306, 5.

[3] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, W., Fan, L., Fougner, C., Gao, L., Gong, C., Hannun, A., Han, T., Johannes, L., Jiang, B., Ju, C., Jun, B., LeGresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng, A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Quan, Z., Raiman, J., Rao, V., Satheesh, S., Seetapun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Z., Wang, Z., Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama, D., Yuan, B., Zhan, J., and Zhu, Z. Deep speech 2 : End-to-end speech recognition in English and Mandarin. In *Proceedings of The 33rd International Conference on Machine Learning* (New York, New York, USA, 20–22 Jun 2016), M. F. Balcan and K. Q. Weinberger, Eds., vol. 48 of *Proceedings of Machine Learning Research*, PMLR, pp. 173–182.

[4] Anderson, G. D., Vamanamurthy, M. K., and Vuorinen, M. Inequalities for quasiconformal mappings in space. *Pacific J. Math. 160*, 1 (1993), 1–18.

[5] Audibert, J.-Y., and Tsybakov, A. B. Fast learning rates for plug-in classifiers. *Ann. Statist. 35*, 2 (2007), 608–633.

[6] Bach, F., and Moulines, E. Non-strongly-convex smooth stochastic approximation with convergence rate O (1/n). In *Advances in Neural Information*

Processing Systems (2013), C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26, Curran Associates, Inc.

[7] Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory 39*, 3 (1993), 930–945.

[8] Barron, A. R. Approximation and estimation bounds for artificial neural networks. *Machine learning 14*, 1 (1994), 115–133.

[9] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc. 101*, 473 (2006), 138–156.

[10] Bartlett, P. L., Montanari, A., and Rakhlin, A. Deep learning: a statistical viewpoint. *Acta Numerica 30* (2021), 87–201.

[11] Bauer, B., and Kohler, M. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Ann. Statist. 47*, 4 (2019), 2261–2285.

[12] Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research 18*, 153 (2018), 1–43.

[13] Baydin, A. G., Pearlmutter, B. A., Syme, D., Wood, F., and Torr, P. Gradients without backpropagation. *arXiv preprint arXiv:2202.08587* (2022).

[14] Bedford, T., and Cooke, R. M. Probability density decomposition for conditionally dependent random variables modeled by vines. *Ann. Math. Artif. Intell. 32*, 1-4 (2001), 245–268.

[15] Belkin, M., Rakhlin, A., and Tsybakov, A. B. Does data interpolation contradict statistical optimality? In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (16–18 Apr 2019), K. Chaudhuri and M. Sugiyama, Eds., vol. 89 of *Proceedings of Machine Learning Research*, PMLR, pp. 1611–1619.

[16] Bennett, G. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association 57*, 297 (1962), 33–45.

[17] Benveniste, A., Métivier, M., and Priouret, P. *Adaptive algorithms and stochastic approximations*, vol. 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1990. Translated from the French by Stephen S. Wilson.

[18] Berend, D., and Tassa, T. Improved bounds on Bell numbers and on moments of sums of random variables. *Probab. Math. Statist. 30*, 2 (2010), 185–205.

[19] Besag, J. Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B 36* (1974), 192–236.

[20] Birgé, L., and Massart, P. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli 4*, 3 (1998), 329–375.

[21] Bishop, C. M. *Pattern recognition and machine learning*. Information Science and Statistics. Springer, New York, 2006.

[22] Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., and Zieba, K. End to end learning for self-driving cars. *arXiv e-prints* (2016), arXiv:1604.07316.

[23] Bos, T., and Schmidt-Hieber, J. Simulation-code: A supervised deep learning method for nonparametric density estimation. `https://github.com/Bostjm/Simulation-code`, Apr. 2023.

[24] Bos, T., and Schmidt-Hieber, J. Simulation code: Convergence guarantees for forward gradient descent in the linear regression model. `https://github.com/Bostjm/SimulationCodeForwardGradient`, Jan. 2024.

[25] Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities*. Oxford University Press, Oxford, 2013.

[26] Brechmann, E. C., Czado, C., and Aas, K. Truncated regular vines in high dimensions with application to financial data. *Canad. J. Statist. 40*, 1 (2012), 68–85.

[27] Breiman, L. Hinging hyperplanes for regression, classification, and function approximation. *IEEE Trans. Inform. Theory 39*, 3 (1993), 999–1013.

[28] Cherubini, U., Luciano, E., and Vecchiato, W. *Copula methods in finance*. Wiley Finance Series. John Wiley & Sons, Ltd., Chichester, 2004.

[29] Choromanska, A., Henaff, M., Mathieu, M., Ben Arous, G., and LeCun, Y. The loss durfaces of multilayer networks. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics* (San Diego, California, USA, 09–12 May 2015), G. Lebanon and S. V. N. Vishwanathan, Eds., vol. 38 of *Proceedings of Machine Learning Research*, PMLR, pp. 192–204.

[30] Cireşan, D., Meier, U., and Schmidhuber, J. Multi-column deep neural networks for image classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), pp. 3642–3649.

[31] Clara, G., Langer, S., and Schmidt-Hieber, J. Dropout regularization versus ℓ_2-penalization in the linear model. *arXiv e-prints* (2023), arXiv:2306.10529.

[32] Conn, A. R., Scheinberg, K., and Vicente, L. N. *Introduction to derivative-free optimization*, vol. 8 of *MPS/SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Programming Society (MPS), Philadelphia, PA, 2009.

[33] Crick, F. The recent excitement about neural networks. *Nature 337* (1989), 129–132.

[34] Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems 2*, 4 (1989), 303–314.

[35] Czado, C. *Analyzing dependent data with vine copulas*, vol. 222 of *Lecture Notes in Statistics*. Springer, Cham, 2019. A practical guide with R.

[36] Czado, C., and Nagler, T. Vine copula based modeling. *Annu. Rev. Stat. Appl. 9* (2022), 453–477.

[37] Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems* (2014), Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27, Curran Associates, Inc.

[38] Devroye, L., Györfi, L., and Lugosi, G. *A probabilistic theory of pattern recognition*, vol. 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

[39] Drouet Mari, D., and Kotz, S. *Correlation and dependence*. Imperial College Press, London; distributed by World Scientific Publishing Co., Inc., River Edge, NJ, 2001.

[40] Dua, D., and Graff, C. UCI machine learning repository, 2017.

[41] Duchi, J. C., Jordan, M. I., Wainwright, M. J., and Wibisono, A. Optimal rates for zero-order convex optimization: the power of two function evaluations. *IEEE Trans. Inform. Theory 61*, 5 (2015), 2788–2806.

[42] Dudley, R. M. A course on empirical processes. In *École d'été de probabilités de Saint-Flour, XII—1982*, vol. 1097 of *Lecture Notes in Math.* Springer, Berlin, 1984, pp. 1–142.

[43] Durante, F., and Sempi, C. Copula theory: an introduction. In *Copula theory and its applications*, vol. 198 of *Lect. Notes Stat. Proc.* Springer, Heidelberg, 2010, pp. 3–31.

[44] Efromovich, S. *Nonparametric curve estimation.* Springer Series in Statistics. Springer-Verlag, New York, 1999.

[45] Efron, B., and Tibshirani, R. Using specially designed exponential families for density estimation. *Ann. Statist. 24*, 6 (1996), 2431–2461.

[46] Funahashi, K.-I. On the approximate realization of continuous mappings by neural networks. *Neural Networks 2*, 3 (1989), 183–192.

[47] Gänssler, P. *Empirical processes*, vol. 3 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series.* Institute of Mathematical Statistics, Hayward, CA, 1983.

[48] Gao, Z., and Hastie, T. LinCDE: conditional density estimation via Lindsey's method. *J. Mach. Learn. Res. 23* (2022), Paper No. [52], 55.

[49] Glorot, X., and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (2010), JMLR Workshop and Conference Proceedings, pp. 249–256.

[50] Glorot, X., Bordes, A., and Bengio, Y. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (Fort Lauderdale, FL, USA, 11–13 Apr 2011), G. Gordon, D. Dunson, and M. Dudík, Eds., vol. 15 of *Proceedings of Machine Learning Research*, PMLR, pp. 315–323.

[51] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning.* MIT Press, 2016. `http://www.deeplearningbook.org`.

[52] Greenspan, H., van Ginneken, B., and Summers, R. M. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging 35*, 5 (2016), 1153–1159.

[53] Grossberg, S. Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science 11*, 1 (1987), 23–63.

[54] Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. *A distribution-free theory of nonparametric regression.* Springer Series in Statistics. Springer-Verlag, New York, 2002.

[55] Györfi, L., and Walk, H. On the averaged stochastic approximation for linear regression. *SIAM J. Control Optim. 34*, 1 (1996), 31–61.

[56] Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning*, second ed. Springer Series in Statistics. Springer, New York, 2009. Data mining, inference, and prediction.

[57] Haussler, D., and Opper, M. Mutual information, metric entropy and cumulative relative entropy risk. *Ann. Statist. 25*, 6 (1997), 2451–2492.

[58] He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 1026–1034.

[59] Heckerman, E., and Nathwani, N. Toward normative expert systems: Part II. Probability-based representations for efficient knowledge acquisition and inference. *Methods of Information in medicine 31*, 02 (1992), 106–116.

[60] Hinton, G. E., Osindero, S., and Teh, Y.-W. A fast learning algorithm for deep belief nets. *Neural Comput. 18*, 7 (2006), 1527–1554.

[61] Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural Networks 2*, 5 (1989), 359–366.

[62] Horowitz, J. L., and Mammen, E. Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *Ann. Statist. 35*, 6 (2007), 2589–2619.

[63] Hsu, D., Kakade, S. M., and Zhang, T. Random design analysis of ridge regression. *Found. Comput. Math. 14*, 3 (2014), 569–600.

[64] Johnson, N. L., and Kotz, S. On some generalized Farlie-Gumbel-Morgenstern distributions. II. Regression, correlation and further generalizations. *Comm. Statist.—Theory Methods A6*, 6 (1977), 485–496.

[65] Jones, L. K. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. Statist. 20*, 1 (1992), 608–613.

[66] Kantorovitz, S. *Several real variables*. Springer Undergraduate Mathematics Series. Springer, [Cham], 2016.

[67] Kim, Y., Ohn, I., and Kim, D. Fast convergence rates of deep neural networks for classification. *Neural Networks 138* (2021), 179–197.

[68] Kingma, D. P., and Welling, M. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning 12*, 4 (2019), 307–392.

[69] Kingman, J. F. C. *Poisson processes*, vol. 3 of *Oxford Studies in Probability*. The Clarendon Press, Oxford University Press, New York, 1993. Oxford Science Publications.

[70] Kirshner, S. Learning with tree-averaged densities and distributions. In *Advances in Neural Information Processing Systems* (2007), J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., vol. 20, Curran Associates, Inc.

[71] Kohler, M., and Krzyżak, A. Adaptive regression estimation with multilayer feedforward neural networks. *J. Nonparametr. Stat. 17*, 8 (2005), 891–913.

[72] Kohler, M., and Krzyżak, A. Nonparametric regression based on hierarchical interaction models. *IEEE Trans. Inform. Theory 63*, 3 (2017), 1620–1630.

[73] Kohler, M., Krzyżak, A., and Langer, S. Estimation of a function of low local dimensionality by deep neural networks. *IEEE Trans. Inform. Theory 68*, 6 (2022), 4032–4042.

[74] Kohler, M., Krzyzak, A., and Walter, B. On the rate of convergence of image classifiers based on convolutional neural networks. *arXiv preprint arXiv:2003.01526* (2020).

[75] Kohler, M., and Langer, S. On the rate of convergence of fully connected very deep neural network regression estimates. *arXiv e-prints* (2019), arXiv:1908.11133.

[76] Kohler, M., and Langer, S. Statistical theory for image classification using deep convolutional neural networks with cross-entropy loss. *arXiv preprint arXiv:2011.13602* (2020).

[77] Kohler, M., and Langer, S. On the rate of convergence of fully connected deep neural network regression estimates. *Ann. Statist. 49*, 4 (2021), 2231–2249.

[78] Koller, D., and Friedman, N. *Probabilistic graphical models*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2009.

[79] Korb, K. B., and Nicholson, A. E. *Bayesian artificial intelligence*. Chapman & Hall/CRC Computer Science and Data Analysis Series. Chapman & Hall/CRC, Boca Raton, FL, 2004.

[80] Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 1097–1105.

[81] Kushner, H. J., and Yin, G. G. *Stochastic approximation and recursive algorithms and applications*, second ed., vol. 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 2003.

[82] Lakshminarayanan, C., and Szepesvari, C. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics* (09–11 Apr 2018), A. Storkey and F. Perez-Cruz, Eds., vol. 84 of *Proceedings of Machine Learning Research*, PMLR, pp. 1347–1355.

[83] Larson, J., Menickelly, M., and Wild, S. M. Derivative-free optimization methods. *Acta Numer. 28* (2019), 287–404.

[84] Launay, J., Poli, I., Boniface, F., and Krzakala, F. Direct feedback alignment scales to modern deep learning tasks and architectures. In *Advances in Neural Information Processing Systems* (2020), H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., pp. 9346–9360.

[85] Lauritzen, S. L. *Graphical models*, vol. 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1996. Oxford Science Publications.

[86] Le Cun, Y. Learning process in an asymmetric threshold network. In *Disordered systems and biological organization*, vol. 20. Springer, 1986, pp. 233–240.

[87] Leibig, C., Allken, V., Ayhan, M. S., Berens, P., and Wahl, S. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports 7*, 1 (2017), 1–14.

[88] Lillicrap, T. P., Cownden, D., Tweed, D. B., and Akerman, C. J. Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications 7*, 1 (2016), 13276.

[89] Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., and Hinton, G. Backpropagation and the brain. *Nature Reviews Neuroscience 21* (2020), 335–346.

[90] Lindsey, J. K. Comparison of probability distributions. *J. Roy. Statist. Soc. Ser. B 36* (1974), 38–47.

[91] Lindsey, J. K. Construction and comparison of statistical models. *J. Roy. Statist. Soc. Ser. B 36* (1974), 418–425.

[92] Liu, S., Chen, P.-Y., Kailkhura, B., Zhang, G., Hero III, A. O., and Varshney, P. K. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine 37*, 5 (2020), 43–54.

[93] Loader, C. *Local regression and likelihood*. Statistics and Computing. Springer-Verlag, New York, 1999.

[94] Mammen, E., and Tsybakov, A. B. Smooth discrimination analysis. *Ann. Statist. 27*, 6 (1999), 1808–1829.

[95] McCulloch, W. S., and Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics 5* (1943), 115–133.

[96] Mörters, P., and Peres, Y. *Brownian motion*, vol. 30 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2010.

[97] Moschopoulos, P., and Staniswalis, J. G. Estimation given conditionals from an exponential family. *Amer. Statist. 48*, 4 (1994), 271–275.

[98] Mourtada, J. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *Ann. Statist. 50*, 4 (2022), 2157–2178.

[99] Murphy, K. P. *Machine Learning: a Probabilistic Perspective*. MIT press, 2012.

[100] Nadaraya, E. A. On estimating regression. *Theory of Probability & Its Applications 9*, 1 (1964), 141–142.

[101] Nagler, T., and Czado, C. Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *J. Multivariate Anal. 151* (2016), 69–89.

[102] Nair, V., and Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (2010), pp. 807–814.

[103] Nakada, R., and Imaizumi, M. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *J. Mach. Learn. Res. 21* (2020), Paper No. 174, 38.

[104] Nelsen, R. B. *An introduction to copulas*, second ed. Springer Series in Statistics. Springer, New York, 2006.

[105] Nesterov, Y., and Spokoiny, V. Random gradient-free minimization of convex functions. *Found. Comput. Math. 17*, 2 (2017), 527–566.

[106] Nøkland, A. Direct feedback alignment provides learning in deep neural networks. In *Advances in Neural Information Processing Systems* (2016), D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc.

[107] Nussbaum, M. Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist. 24*, 6 (1996), 2399–2430.

[108] Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Statist. 33* (1962), 1065–1076.

[109] Pearl, J. *Causality*, second ed. Cambridge University Press, Cambridge, 2009.

[110] Petersen, P., and Voigtlaender, F. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks 108* (2018), 296–330.

[111] Petersen, P., and Voigtlaender, F. Optimal learning of high-dimensional classification problems using deep neural networks. *arXiv preprint arXiv:2112.12555* (2021).

[112] Pinelis, I. L'Hospital type results for monotonicity, with applications. *JIPAM. J. Inequal. Pure Appl. Math. 3*, 1 (2002), Article 5, 5.

[113] Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., and Liao, Q. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. *International Journal of Automation and Computing 14*, 5 (2017), 503–519.

[114] Polyak, B. T., and Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim. 30*, 4 (1992), 838–855.

[115] Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647* (2011).

[116] Ray, K., and Schmidt-Hieber, J. The Le Cam distance between density estimation, Poisson processes and Gaussian white noise. *Math. Stat. Learn. 1*, 2 (2018), 101–170.

[117] Ren, M., Kornblith, S., Liao, R., and Hinton, G. Scaling forward gradient with local losses. *arXiv preprint arXiv:2210.03310* (2022).

[118] Resnick, S. *Adventures in stochastic processes*. Birkhäuser Boston, Inc., Boston, MA, 1992.

[119] Rice, J. A. *Mathematical Statistics and Data Analysis (Third Edition)*. Brooks/Cole,Cengage Learning, 2007.

[120] Robbins, H. A remark on Stirling's formula. *Amer. Math. Monthly 62* (1955), 26–29.

[121] Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review 65*, 6 (1958), 386–408.

[122] Rosenblatt, F. *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Spartan books Washington, DC, 1962.

[123] Rosenblatt, M. Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist. 27* (1956), 832–837.

[124] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *Nature 323* (1986), 533–536.

[125] Saxe, A. M., McClelland, J. L., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *ICLR* (2014).

[126] Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks 61* (2015), 85–117.

[127] Schmidt-Hieber, J. Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Statist. 48*, 4 (2020), 1875–1897.

[128] Schmidt-Hieber, J. Interpreting learning in biological neural networks as zero-order optimization method. *arXiv preprint arXiv:2301.11777* (2023).

[129] Schmidt-Hieber, J., and Koolen, W. Hebbian learning inspired estimation of the linear regression parameters from queries. *arXiv preprint* (2023).

[130] Schmidt-Hieber, J., and Zamolodtchikov, P. Local convergence rates of the least squares estimator with applications to transfer learning. *arXiv e-prints* (2022), arXiv:2204.05003.

[131] Scott, D. W. *Multivariate density estimation*, second ed. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, 2015. Theory, practice, and visualization.

[132] Shaffer, J. P. The Gauss-Markov theorem and random regressors. *Amer. Statist. 45*, 4 (1991), 269–273.

[133] Shamir, O. On the complexity of bandit and derivative-free stochastic convex optimization. In *Proceedings of the 26th Annual Conference on Learning Theory* (Princeton, NJ, USA, 12–14 Jun 2013), S. Shalev-Shwartz and I. Steinwart, Eds., vol. 30 of *Proceedings of Machine Learning Research*, PMLR, pp. 3–24.

[134] Shen, G., Jiao, Y., Lin, Y., and Huang, J. Non-asymptotic excess risk bounds for classification with deep convolutional neural networks. *arXiv preprint arXiv:2105.00292* (2021).

[135] Spall, J. C. *Introduction to stochastic search and optimization.* Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2003. Estimation, simulation, and control.

[136] Stöber, J., Joe, H., and Czado, C. Simplified pair copula constructions—limitations and extensions. *J. Multivariate Anal. 119* (2013), 101–118.

[137] Stone, C. J. Optimal rates of convergence for nonparametric estimators. *Ann. Statist. 8*, 6 (1980), 1348–1360.

[138] Stone, C. J. Optimal global rates of convergence for nonparametric regression. *Ann. Statist. 10*, 4 (1982), 1040–1053.

[139] Stone, C. J. Additive regression and other nonparametric models. *Ann. Statist. 13*, 2 (1985), 689–705.

[140] Tarigan, B., and van de Geer, S. A. A moment bound for multi-hinge classifiers. *J. Mach. Learn. Res. 9* (2008), 2171–2185.

[141] Telgarsky, M. Benefits of depth in neural networks. In *29th Annual Conference on Learning Theory* (Columbia University, New York, New York, USA, 23–26 Jun 2016), V. Feldman, A. Rakhlin, and O. Shamir, Eds., vol. 49 of *Proceedings of Machine Learning Research*, PMLR, pp. 1517–1539.

[142] Trappenberg, T. P. *Fundamentals of Computational Neuroscience: Third Edition.* Oxford University Press, 12 2022.

[143] Triantafyllopoulos, K. On the central moments of the multidimensional Gaussian distribution. *Math. Sci. 28*, 2 (2003), 125–128.

[144] Tsybakov, A. B. Optimal rates of aggregation. In *Learning Theory and Kernel Machines* (Berlin, Heidelberg, 2003), B. Schölkopf and M. K. Warmuth, Eds., Springer Berlin Heidelberg, pp. 303–313.

[145] Tsybakov, A. B. Optimal aggregation of classifiers in statistical learning. *Ann. Statist. 32*, 1 (2004), 135–166.

[146] Tsybakov, A. B. *Introduction to Nonparametric Estimation.* Springer Series in Statistics. Springer New York, 2008.

[147] van de Geer, S. A. *Empirical Processes in M-Estimation.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

[148] van der Vaart, A. W., and Wellner, J. A. *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[149] van Erven, T., and Harremoës, P. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inform. Theory 60*, 7 (2014), 3797–3820.

[150] Vapnik, V. N. *The nature of statistical learning theory*, second ed. Statistics for Engineering and Information Science. Springer-Verlag, New York, 2000.

[151] Wand, M. P., and Jones, M. C. *Kernel smoothing*, vol. 60 of *Monographs on Statistics and Applied Probability.* Chapman and Hall, Ltd., London, 1995.

[152] Wasserman, L. *All of statistics.* Springer Texts in Statistics. Springer-Verlag, New York, 2004.

[153] Wasserman, L. *All of nonparametric statistics.* Springer Texts in Statistics. Springer, New York, 2006.

[154] Watson, G. S. Smooth regression analysis. *Sankhyā Ser. A 26* (1964), 359–372.

[155] Whittington, J. C. R., and Bogacz, R. An approximation of the error backpropagation algorithm in a predictive coding network with local Hebbian synaptic plasticity. *Neural Comput. 29*, 5 (2017), 1229–1262.

[156] Widrow, B., and Hoff, M. E. Adaptive switching circuits. In *IRE WESCON convention record* (1960), vol. 4, New York, pp. 96–104.

[157] Wong, W. H., and Severini, T. A. On maximum likelihood estimation in infinite-dimensional parameter spaces. *Ann. Statist. 19*, 2 (1991), 603–632.

[158] Wong, W. H., and Shen, X. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist. 23*, 2 (1995), 339–362.

[159] Yang, Y., and Barron, A. Information-theoretic determination of minimax rates of convergence. *Ann. Statist. 27*, 5 (1999), 1564–1599.

[160] Yarotsky, D. Error bounds for approximations with deep ReLU networks. *Neural Networks 94* (2017), 103–114.