References

[1] P. A. M. Dirac, *The Principles of Quantum Mechanics*. No. 27 in The International Series of Monographs on Physics. Oxford University Press, 1930.

[2] J. von Neumann, *Mathematical Foundations of Quantum Mechanics*. Princeton University Press, 1955.

[3] S. Carnot, *Reflections on the Motive Power of Heat and on Machines Fitted to Develop that Power*. J. Wiley, 1890. Trans. by R. H. Thurston from *Réflexions sur la puissance motrice du feu et sur les machines propres à développer cette puissance* (1824).

[4] M. Bessarab, *Landau*. M., Moscow worker, 1971. Trans. by B. Hanin from the original Russian source. www.ega-math.narod.ru/Landau/Dau1971.htm.

[5] J. Polchinski, "Memories of a Theoretical Physicist," arXiv:1708.09093 [physics.hist-ph].

[6] F. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanism*, tech. rep., Cornell Aeronautical Lab, Inc., 1961.

[7] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics* **5** no. 4, (1943) 115–133.

[8] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review* **65** no. 6, (1958) 386.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing System* **25**, (2012) 1097–1105.

[10] K. Fukushima, "Neocognitron: A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics* **36** no. 4, (1980) 193–202.

[11] Y. LeCun, Generalization and Network Design Strategies, tech. rep. CRG-TR-89-4, Department of Computer Science, University of Toronto, 1989.

[12] Y. LeCun, B. Boser, J. S. Denker, et al., "Backpropagation applied to handwritten zip code recognition," *Neural Computation* **1** no. 4, (1989) 541–551.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE* **86** no. 11, (1998) 2278–2324.

[14] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in *Advances in Neural Information Processing Systems* **30**, (2017) 5998–6008. arXiv:1706.03762 [cs.CL].

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure*

Published online by Cambridge University Press

of Cognition. Volume 1: Foundations, D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, eds., Ch. 8, MIT Press, 1986, pp. 318–362.

[16] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural Networks: Tricks of the Trade*, pp. 9–48. Springer, 1998.

[17] Gallant and White, "There exists a neural network that does not make avoidable mistakes," in *IEEE 1988 International Conference on Neural Networks*, pp. 657–664 vol.1. 1988.

[18] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *International Conference on Machine Learning*. 2010.

[19] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 315–323, JMLR Workshop and Conference Proceedings. 2011.

[20] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *ICML Workshop on Deep Learning for Audio, Speech, and Language Processing*. 2013.

[21] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia, "Incorporating second-order functional knowledge for better option pricing," in *Advances in Neural Information Processing Systems* **13**, (2000) 472–478.

[22] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," arXiv:1710.05941 [cs.NE].

[23] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," arXiv:1606.08415 [cs.LG].

[24] A. M. Turing, "The chemical basis of morphogenesis," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **237** no. 641, (1952) 37–72.

[25] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," arXiv:1312.6120 [cs.NE].

[26] J. A. Zavatone-Veth and C. Pehlevan, "Exact priors of finite neural networks," arXiv:2104.11734 [cs.LG].

[27] J. McGreevy, "Holographic duality with a view toward many-body physics," *Advances in High Energy Physics* **2010** (2010) 723105, arXiv:0909.0518 [hep-th].

[28] R. M. Neal, "Priors for infinite networks," in *Bayesian Learning for Neural Networks*, pp. 29–53. Springer, 1996.

[29] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein, "Deep neural networks as Gaussian processes," in *International Conference on Learning Representations*. 2018. arXiv:1711.00165 [stat.ML].

[30] A. G. de G. Matthews, M. Rowland, J. Hron, R. E. Turner, and Z. Ghahramani, "Gaussian process behaviour in wide deep neural networks," in *International Conference on Learning Representations*. 2018. arXiv:1804.11271 [stat.ML].

[31] S. Yaida, "Non-Gaussian processes and neural networks at finite widths," in *Mathematical and Scientific Machine Learning Conference*. 2020. arXiv:1910.00019 [stat.ML].

[32] F. J. Dyson, "The *S* matrix in quantum electrodynamics," *Physical Review* **75** (Jun, 1949) 1736–1755.

[33] J. Schwinger, "On the Green's functions of quantized fields. I," *Proceedings of the National Academy of Sciences* **37** no. 7, (1951) 452–455.

Published online by Cambridge University Press

[34] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014*, (2014) 818–833.

[35] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in Neural Information Processing Systems* **20**, (2008) 1177–1184.

[36] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv:1810.04805 [cs.CL].

[37] M. Gell-Mann and F. E. Low, "Quantum electrodynamics at small distances," *Physical Review* **95** (Sep, 1954) 1300–1312.

[38] K. G. Wilson, "Renormalization group and critical phenomena. I. Renormalization group and the Kadanoff scaling picture," *Physical Review B* **4** (Nov, 1971) 3174–3183.

[39] K. G. Wilson, "Renormalization group and critical phenomena. II. Phase-space cell analysis of critical behavior," *Physical Review B* **4** (Nov, 1971) 3184–3205.

[40] E. C. G. Stueckelberg de Breidenbach and A. Petermann, "Normalization of constants in the quanta theory," *Helvetica Physica Acta* **26** (1953) 499–520.

[41] N. Goldenfeld, *Lectures on Phase Transitions and the Renormalization Group*. CRC Press, 2018.

[42] J. Cardy, *Scaling and Renormalization in Statistical Physics*. Cambridge Lecture Notes in Physics. Cambridge University Press, 1996.

[43] M. Minsky and S. A. Papert, *Perceptrons: An Introduction to Computational Geometry*. MIT Press, 1988.

[44] S. Coleman, *Aspects of Symmetry: Selected Erice Lectures*. Cambridge University Press, 1985.

[45] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, "Exponential expressivity in deep neural networks through transient chaos," in *Advances in Neural Information Processing Systems* **29**, (2016) 3360–3368. arXiv:1606.05340 [stat.ML].

[46] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, "On the expressive power of deep neural networks," in *International Conference on Machine Learning*, pp. 2847–2854. 2017. arXiv:1606.05336 [stat.ML].

[47] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, "Deep information propagation," in *5th International Conference on Learning Representations*. 2017. arXiv:1611.01232 [stat.ML].

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034. 2015. arXiv:1502.01852 [cs.CV].

[49] L. Kadanoff, "Critical behavior. Universality and scaling," in *Proceedings of the International School of Physics Enrico Fermi, Course LI (27 July - 8 August 1970)*. 1971.

[50] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[51] L. Froidmont, *On Christian Philosophy of the Soul*. 1649.

[52] D. J. MacKay, "Probable networks and plausible predictions–a review of practical Bayesian methods for supervised neural networks," *Network: Computation in Neural Systems* **6** no. 3, (1995) 469–505.

[53] C. K. I. Williams, "Computing with infinite networks," in *Advances in Neural Information Processing Systems* **9**, (1996) 295–301.

[54] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*. 2015. arXiv:1503.02531 [stat.ML].

Published online by Cambridge University Press

[55] D. Hebb, *The Organization of Behavior: A Neuropsychological Theory*. Taylor & Francis, 2005.

[56] S. Coleman, "Sidney Coleman's Dirac lecture 'quantum mechanics in your face'," arXiv:2011.12671 [physics.hist-ph].

[57] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Advances in Neural Information Processing Systems* **31**, (2018) 8571–8580. arXiv:1806.07572 [cs.LG].

[58] S. Hochreiter, *Untersuchungen zu dynamischen neuronalen Netzen*. Diploma, Technische Universität München, 1991.

[59] Y. Bengio, P. Frasconi, and P. Simard, "The problem of learning long-term dependencies in recurrent networks," in *IEEE International Conference on Neural Networks*, vol. 3, pp. 1183–1188, IEEE. 1993.

[60] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, pp. 1310–1318, PMLR. 2013. arXiv:1211.5063 [cs.LG].

[61] M. Kline, *Mathematical Thought From Ancient to Modern Times: Volume 3*. Oxford University Press, 1990.

[62] J. Lee, L. Xiao, S. Schoenholz, et al., "Wide neural networks of any depth evolve as linear models under gradient descent," in *Advances in Neural Information Processing Systems* **32**, (2019) 8572–8583. arXiv:1902.06720 [stat.ML].

[63] E. Fix and J. Hodges, "Discriminatory analysis. Nonparametric discrimination: Consistency properties," *USAF School of Aviation Medicine, Project Number: 21-49-004, Report Number: 4* (1951) .

[64] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory* **13** (1967) 21–27.

[65] A. Einstein, "On the method of theoretical physics," *Philosophy of Science* **1** no. 2, (1934) 163–169.

[66] B. Hanin and M. Nica, "Finite depth and width corrections to the neural tangent kernel," in *International Conference on Learning Representations*. 2020. arXiv:1909.05989 [cs.LG].

[67] E. Dyer and G. Gur-Ari, "Asymptotics of wide networks from Feynman diagrams," in *International Conference on Learning Representations*. 2020. arXiv:1909.11304 [cs.LG].

[68] G. F. Giudice, "Naturally speaking: The naturalness criterion and physics at the LHC," arXiv:0801.2562 [hep-ph].

[69] F. J. Dyson, "Forward," in *Classic Feynman: All the Adventures of a Curious Character*, R. Leighton, ed., W. W. Norton & Company Ltd., 2006, pp. 5–9.

[70] L. Chizat, E. Oyallon, and F. Bach, "On lazy training in differentiable programming," in *Advances in Neural Information Processing Systems* **32**, (2019) 2937–2947. arXiv:1812.07956 [math.OC].

[71] D. J. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[72] J. Kaplan, S. McCandlish, T. Henighan, et al., "Scaling laws for neural language models," arXiv:2001.08361 [cs.LG].

[73] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," arXiv:1611.03530 [cs.LG].

[74] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural Computation* **8** no. 7, (1996) 1341–1390.

Published online by Cambridge University Press

[75] D. H. Wolpert and W. G. Macready, “No free lunch theorems for optimization,” *IEEE Transactions on Evolutionary Computation* **1** no. 1, (1997) 67–82.

[76] L. Boltzmann, “On certain questions of the theory of gases,” *Nature* **51** no. 1322, (1895) 413–415.

[77] L. Boltzmann, *Lectures on Gas Theory*. Berkeley, University of California Press, 1964. Trans. by S. G. Brush from *Vorlesungen ueber Gastheorie* (2 vols., 1896 & 1898).

[78] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal* **27** no. 3, (1948) 379–423.

[79] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal* **27** no. 4, (1948) 623–656.

[80] E. T. Jaynes, “Information theory and statistical mechanics,” *Physical Review* **106** (May, 1957) 620–630.

[81] E. T. Jaynes, “Information theory and statistical mechanics. II,” *Physical Review* **108** (Oct, 1957) 171–190.

[82] Y. LeCun, J. Denker, and S. Solla, “Optimal brain damage,” in *Advances in Neural Information Processing Systems* **2**, (1990) 598–605.

[83] J. Frankle and M. Carbin, “The lottery ticket hypothesis: Finding sparse, trainable neural networks,” in *International Conference on Learning Representations*. 2019. arXiv:1803.03635 [cs.LG].

[84] T. Banks and A. Zaks, “On the phase structure of vector-like gauge theories with massless fermions,” *Nuclear Physics B* **196** no. 2, (1982) 189–204.

[85] R. Linsker, “Self-organization in a perceptual network,” *Computer* **21** no. 3, (1988) 105–117.

[86] S. Becker and G. E. Hinton, “Self-organizing neural network that discovers surfaces in random-dot stereograms,” *Nature* **355** no. 6356, (1992) 161–163.

[87] B. Gale, R. Zemeckis, M. J. Fox, and C. Lloyd, *Back to the Future Part II*. Universal Pictures, 1989.

[88] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778. 2016.

[89] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International Conference on Machine Learning*, pp. 448–456. 2015. arXiv:1502.03167 [cs.LG].

[90] A. Veit, M. Wilber, and S. Belongie, “Residual networks behave like ensembles of relatively shallow networks,” in *Advances in Neural Information Processing Systems* **30**, (2016) 550–558. arXiv:1605.06431 [cs.CV].

[91] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” in *Deep Learning Symposium, Neural Information Processing Systems*. 2016. arXiv:1607.06450 [stat.ML].

Published online by Cambridge University Press

Published online by Cambridge University Press