参考文献 Bibliography

[AAB21] A. Agrawal, A. Ali, and S. Boyd. “Minimum-distortion embedding”. en. In: *Foundations and Trends in Machine Learning* 14.3 (2021), pp. 211–378.

[AB08] C. Archambeau and F. Bach. “Sparse probabilistic projections”. In: *NIPS*. 2008.

[AB14] G. Alain and Y. Bengio. “What Regularized Auto-Encoders Learn from the Data-Generating Distribution”. In: *JMLR* (2014).

[AC16] D. K. Agarwal and B.-C. Chen. *Statistical Methods for Recommender Systems*. en. 1st edition. Cambridge University Press, 2016.

[Ace] “The Turing Test is Bad for Business”. In: (2021).

[AEH+18] S. Abu-El-Haija, B. Perozzi, R. Al-Rfou, and A. A. Alemi. “Watch your step: Learning node embeddings via graph attention”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 9180–9190.

[AEHPAR17] S. Abu-El-Haija, B. Perozzi, and R. Al-Rfou. “Learning Edge Representations via Low-Rank Asymmetric Projections”. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. CIKM ’17. 2017, 1787–1796.

[AEM18] Ö. D. Akyildiz, V. Elvira, and J. Miguez. “The Incremental Proximal Method: A Probabilistic Perspective”. In: *ICASSP*. 2018.

[AFF19] C. Aicher, N. J. Foti, and E. B. Fox. “Adaptively Truncating Backpropagation Through Time to Control Gradient Bias”. In: (2019). arXiv: 1905.07473 [cs.LG].

[Agg16] C. C. Aggarwal. *Recommender Systems: The Textbook*. en. 1st ed. 2016 edition. Springer, 2016.

[Agg20] C. C. Aggarwal. *Linear Algebra and Optimization for Machine Learning: A Textbook*. en. 1st ed. 2020 edition. Springer, 2020.

[AGM19] V. Amrhein, S. Greenland, and B. McShane. “Scientists rise up against statistical significance”. In: *Nature* 567.7748 (2019), p. 305.

[Agr70] A. Agrawala. “Learning with a probabilistic teacher”. In: *IEEE Transactions on Information Theory* 16.4 (1970), pp. 373–379.

[AH19] C. Allen and T. Hospedales. “Analogies Explained: Towards Understanding Word Embeddings”. In: *ICML*. 2019.

[AHK12] A. Anandkumar, D. Hsu, and S. M. Kakade. “A Method of Moments for Mixture Models and Hidden Markov Models”. In: *COLT*. Vol. 23. Proceedings of Machine Learning Research. PMLR, 2012, pp. 33.1–33.34.

[Ahm+13] A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A. J. Smola. “Distributed large-scale natural graph factorization”. In: *Proceedings of the 22nd international conference on World Wide Web*. ACM. 2013, pp. 37–48.

[AK15] J. Andreas and D. Klein. “When and why are log-linear models self-normalizing?” In: *Proc. ACL*. Association for Computational Linguistics, 2015, pp. 244–249.

[Aka74] H. Akaike. “A new look at the statistical model identification”. In: *IEEE Trans. on Automatic Control* 19.6 (1974).

[AKA91] D. W. Aha, D. Kibler, and M. K. Albert. “Instance-based learning algorithms”. In: *Mach. Learn.* 6.1 (1991), pp. 37–66.

[Aky+19] Ö. D. Akyildiz, É. Chouzenoux, V. Elvira, and J. Míguez. “A probabilistic incremental proximal gradient method”. In: *IEEE Signal Process. Lett.* 26.8 (2019).

[Ala18] J. Alammar. *Illustrated Transformer*. Tech. rep. 2018.

[Alb+17] M. Alber, P.-J. Kindermans, K. Schütt, K.-R. Müller, and F. Sha. “An Empirical Study on The Properties of Random Bases for Kernel Methods”. In: *NIPS*. Curran Associates, Inc., 2017, pp. 2763–2774.

[Alb+18] D. Albanese, S. Riccadonna, C. Donati, and P. Franceschi. “A practical tool for maximal information coefficient anal-

ysis". en. In: *Gigascience* 7.4 (2018), pp. 1–8.

[ALL18] S. Arora, Z. Li, and K. Lyu. "Theoretical Analysis of Auto Rate-Tuning by Batch Normalization". In: (2018). arXiv: 1812.03981 [cs.LG].

[Alm87] L. B. Almeida. "A learning rule for asynchronous perceptrons with feedback in a combinatorial environment." In: *Proceedings, 1st First International Conference on Neural Networks*. Vol. 2. IEEE. 1987, pp. 609–618.

[Alo+09] D. Aloise, A. Deshpande, P. Hansen, and P. Popat. "NP-hardness of Euclidean sum-of-squares clustering". In: *Machine Learning* 75 (2009), pp. 245–249.

[Alp04] E. Alpaydin. *Introduction to machine learning*. MIT Press, 2004.

[Ami+19] E. Amid, M. K. Warmuth, R. Anil, and T. Koren. "Robust Bi-Tempered Logistic Loss Based on Bregman Divergences". In: *NIPS*. 2019.

[Amo+16] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. "Concrete Problems in AI Safety". In: (2016). arXiv: 1606.06565 [cs.AI].

[Amo17] Amoeba. *What is the difference between ZCA whitening and PCA whitening*. Stackexchange. 2017.

[And01] C. A. Anderson. "Heat and Violence". In: *Current Directions in Psychological Science* 10.1 (2001), pp. 33–38.

[Ani+20] R. Anil, V. Gupta, T. Koren, K. Regan, and Y. Singer. "Scalable Second Order Optimization for Deep Learning". In: (2020). arXiv: 2002.09018 [cs.LG].

[Ans73] F. J. Anscombe. "Graphs in Statistical Analysis". In: *Am. Stat.* 27.1 (1973), pp. 17–21.

[AO03] J.-H. Ahn and J.-H. Oh. "A Constrained EM Algorithm for Principal Component Analysis". In: *Neural Computation* 15 (2003), pp. 57–65.

[Arc+19] F. Arcadu, F. Benmansour, A. Maunz, J. Willis, Z. Haskova, and M. Prunotto. "Deep learning algorithm predicts diabetic retinopathy progression in individual patients". en. In: *NPJ Digit Med* 2 (2019), p. 92.

[Ard+20] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber. "Common Voice: A Massively-Multilingual Speech Corpus". In: *Proceedings of The 12th Language Resources and Evaluation Conference*. 2020, pp. 4218–4222.

[Arj21] M. Arjovsky. "Out of Distribution Generalization in Machine Learning". In: (2021). arXiv: 2103.02667 [stat.ML].

[Arn+19] S. M. R. Arnold, P.-A. Manzagol, R. Babanezhad, I. Mitliagkas, and N. Le Roux. "Reducing the variance in online optimization by transporting past gradients". In: *NIPS*. 2019.

[Aro+16] S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski. "A Latent Variable Model Approach to PMI-based Word Embeddings". In: *TACL* 4 (2016), pp. 385–399.

[Aro+19] L. Aroyo, A. Dumitrache, O. Inel, Z. Szlávik, B. Timmermans, and C. Welty. "Crowdsourcing Inclusivity: Dealing with Diversity of Opinions, Perspectives and Ambiguity in Annotated Data". In: *WWW*. WWW '19. Association for Computing Machinery, 2019, pp. 1294–1295.

[Aro+21] R. Arora et al. *Theory of deep learning*. 2021.

[ARZP19] R. Al-Rfou, D. Zelle, and B. Perozzi. "DDGK: Learning Graph Representations for Deep Divergence Graph Kernels". In: *Proceedings of the 2019 World Wide Web Conference on World Wide Web* (2019).

[AS17] A. Achille and S. Soatto. "On the Emergence of Invariance and Disentangling in Deep Representations". In: (2017). arXiv: 1706.01350 [cs.LG].

[AS19] A. Achille and S. Soatto. "Where is the Information in a Deep Neural Network?" In: (2019). arXiv: 1905.12213 [cs.LG].

[Ash18] J. Asher. "A Rise in Murder? Let's Talk About the Weather". In: *The New York Times* (2018).

[ASR15] A. Ali, S. M. Shamsuddin, and A. L. Ralescu. "Classification with class imbalance problem: A Review". In: *Int. J. Advance Soft Compu. Appl* 7.3 (2015).

[Ath+19] B. Athiwaratkun, M. Finzi, P. Izmailov, and A. G. Wilson. "There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average". In: *ICLR*. 2019.

[AV07] D. Arthur and S. Vassilvitskii. "k-means++: the advantages of careful seeding". In: *Proc. 18th ACM-SIAM symp. on Discrete algorithms*. 2007, 1027–1035.

[AWS19] E. Amid, M. K. Warmuth, and S. Srinivasan. "Two-temperature logistic regression based on the Tsallis divergence". In: *AISTATS*. 2019.

[Axl15] S. Axler. *Linear algebra done right*. 2015.

[BA10] R. Bailey and J. Addison. *A Smoothed-Distribution Form of Nadaraya-Watson Estimation*. Tech. rep. 10-30. Univ. Birmingham, 2010.

[BA97a] A. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis*. Oxford, 1997.

[BA97b] L. A. Breslow and D. W. Aha. "Simplifying decision trees: A survey". In: *Knowl. Eng. Rev.* 12.1 (1997), pp. 1–40.

[Bab19] S. Babu. *A 2019 guide to Human Pose Estimation with Deep Learning*. 2019.

[Bac+16] O. Bachem, M. Lucic, H. Hassani, and A. Krause. "Fast and Provably Good Seedings for k-Means". In: *NIPS*. 2016, pp. 55–63.

[Bah+12] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii. "Scalable k-Means++". In: *VLDB*. 2012.

[Bah+20] Y. Bahri, J. Kadmon, J. Pennington, S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli. "Statistical Mechanics of Deep Learning". In: *Annu. Rev. Condens. Matter Phys.* (2020).

[BAP14] P. Bachman, O. Alsharif, and D. Precup. "Learning with pseudo-ensembles". In: *Advances in neural information processing systems*. 2014, pp. 3365–3373.

[Bar09] M. Bar. "The proactive brain: memory for predictions". en. In: *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364.1521 (2009), pp. 1235–1243.

[Bar19] J. T. Barron. "A General and Adaptive Robust Loss Function". In: *CVPR*. 2019.

[Bat+18] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, et al. "Relational inductive biases, deep learning, and graph networks". In: *arXiv preprint arXiv:1806.01261* (2018).

[BB08] O. Bousquet and L. Bottou. "The Tradeoffs of Large Scale Learning". In: *NIPS*. 2008, pp. 161–168.

[BB11] L. Bottou and O. Bousquet. "The Tradeoffs of Large Scale Learning". In: *Optimization for Machine Learning*. Ed. by S. Sra, S. Nowozin, and S. J. Wright. MIT Press, 2011, pp. 351–368.

[BBV11] R. Benassi, J. Bect, and E. Vazquez. "Bayesian optimization using sequential Monte Carlo". In: (2011). arXiv: 1111.4802 [math.OC].

[BC17] D. Beck and T. Cohn. "Learning Kernels over Strings using Gaussian Processes". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Vol. 2. 2017, pp. 67–73.

[BCB15] D. Bahdanau, K. Cho, and Y. Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *ICLR*. 2015.

[BCD01] L. Brown, T. Cai, and A. DasGupta. "Interval Estimation for a Binomial Proportion". In: *Statistical Science* 16.2 (2001), pp. 101–133.

[BCN18] L. Bottou, F. E. Curtis, and J. Nocedal. "Optimization Methods for Large-Scale Machine Learning". In: *SIAM Rev.* 60.2 (2018), pp. 223–311.

[BCV13] Y. Bengio, A. Courville, and P. Vincent. "Representation learning: a review and new perspectives". en. In: *IEEE PAMI* 35.8 (2013), pp. 1798–1828.

[BD20] B. Barz and J. Denzler. "Do We Train on Test Data? Purging CIFAR of Near-Duplicates". In: *J. of Imaging* 6.6 (2020).

[BD21] D. G. T. Barrett and B. Dherin. "Implicit Gradient Regularization". In: *ICLR*. 2021.

[BD87] G. Box and N. Draper. *Empirical Model-Building and Response Surfaces*. Wiley, 1987.

[BDEL03] S. Ben-David, N. Eiron, and P. M. Long. "On the difficulty of approximately maximizing agreements". In: *J. Comput. System Sci.* 66.3 (2003), pp. 496–514.

[Bel+19] M. Belkin, D. Hsu, S. Ma, and S. Mandal. "Reconciling modern machine-learning practice and the classical bias-variance trade-off". en. In: *PNAS* 116.32 (2019), pp. 15849–15854.

[Ben+04a] Y. Bengio, O. Delalleau, N. Roux, J. Paiement, P. Vincent, and M. Ouimet. "Learning eigenfunctions links spectral embedding and kernel PCA". In: *Neural Computation* 16 (2004), pp. 2197–2219.

[Ben+04b] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. L. Roux, and M. Ouimet. "Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering". In: *NIPS*. MIT Press, 2004, pp. 177–184.

[Ben+15a] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. "Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks". In: *NIPS*. 2015.

[Ben+15b] Y. Bengio, D.-H. Lee, J. Bornschein, T. Mesnard, and Z. Lin. "Towards Biologically Plausible Deep Learning". In: (2015). arXiv: 1502.04156 [cs.LG].

[Ben+17] A. Benavoli, G. Corani, J. Demsar, and M. Zaffalon. "Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis". In: *JMLR* (2017).

[Ber15] D. Bertsekas. *Convex Optimization Algorithms*. Athena Scientific, 2015.

[Ber16] D. Bertsekas. *Nonlinear Programming*. Third. Athena Scientific, 2016.

[Ber+19a] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and

C. Raffel. “Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring”. In: *arXiv preprint arXiv:1911.09785* (2019).

[Ber+19b] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel. “Mixmatch: A holistic approach to semi-supervised learning”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 5049–5059.

[Ber+21] J. Berner, P. Grohs, G. Kutyniok, and P. Petersen. “The Modern Mathematics of Deep Learning”. In: (2021). arXiv: 2105.04026 [cs.LG].

[Ber85] J. Berger. “Bayesian Salesmanship”. In: *Bayesian Inference and Decision Techniques with Applications: Essays in Honor of Bruno deFinetti*. Ed. by P. K. Goel and A. Zellner. North-Holland, 1985.

[Ber99] D. Bertsekas. *Nonlinear Programming*. Second. Athena Scientific, 1999.

[Bey+19] M. Beyeler, E. L. Rounds, K. D. Carlson, N. Dutt, and J. L. Krichmar. “Neural correlates of sparse coding and dimensionality reduction”. en. In: *PLoS Comput. Biol.* 15.6 (2019), e1006908.

[Bey+20] L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, and A. van den Oord. “Are we done with ImageNet?” In: (2020). arXiv: 2006.07159 [cs.CV].

[BFO84] L. Breiman, J. Friedman, and R. Olshen. *Classification and regression trees*. Wadsworth, 1984.

[BG11] P. Buhlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methodology, Theory and Applications*. Springer, 2011.

[BH07] P. Buhlmann and T. Hothorn. “Boosting Algorithms: Regularization, Prediction and Model Fitting”. In: *Statistical Science* 22.4 (2007), pp. 477–505.

[BH69] A. Bryson and Y.-C. Ho. *Applied optimal control: optimization, estimation, and control*. Blaisdell Publishing Company, 1969.

[BH86] J. Barnes and P. Hut. “A hierarchical O(N log N) force-calculation algorithm”. In: *Nature* 324.6096 (1986), pp. 446–449.

[BH89] P. Baldi and K. Hornik. “Neural networks and principal components analysis: Learning from examples without local minima”. In: *Neural Networks* 2 (1989), pp. 53–58.

[Bha+19] A. Bhadra, J. Datta, N. G. Polson, and B. T. Willard. “Lasso Meets Horseshoe: a survey”. In: *Bayesian Anal.* 34.3 (2019), pp. 405–427.

[BHM92] J. S. Bridle, A. J. Heading, and D. J. MacKay. “Unsupervised Classifiers, Mutual Information and’Phantom Targets”. In: *Advances in neural information processing systems*. 1992, pp. 1096–1101.

[BI19] P. Barham and M. Isard. “Machine Learning Systems are Stuck in a Rut”. In: *Proceedings of the Workshop on Hot Topics in Operating Systems*. HotOS ’19. Association for Computing Machinery, 2019, pp. 177–183.

[Bis06] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[Bis94] C. M. Bishop. *Mixture Density Networks*. Tech. rep. NCRG 4288. Neural Computing Research Group, Department of Computer Science, Aston University, 1994.

[Bis99] C. Bishop. “Bayesian PCA”. In: *NIPS*. 1999.

[BJ05] F. Bach and M. Jordan. *A probabilistic interpretation of canonical correlation analysis*. Tech. rep. 688. U. C. Berkeley, 2005.

[BJM06] P. Bartlett, M. Jordan, and J. McAuliffe. “Convexity, Classification, and Risk Bounds”. In: *JASA* 101.473 (2006), pp. 138–156.

[BK07] R. M. Bell and Y. Koren. “Lessons from the Netflix Prize Challenge”. In: *SIGKDD Explor. Newsl.* 9.2 (2007), pp. 75–79.

[BKC17] V. Badrinarayanan, A. Kendall, and R. Cipolla. “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation”. In: *IEEE PAMI* 39.12 (2017).

[BKH16] J. L. Ba, J. R. Kiros, and G. E. Hinton. “Layer Normalization”. In: (2016). arXiv: 1607.06450 [stat.ML].

[BKL10] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. 2010.

[BL04] P. Bickel and E. Levina. “Some theory for Fisher’s linear discriminant function, "Naive Bayes", and some alternatives when there are many more variables than observations”. In: *Bernoulli* 10 (2004), pp. 989–1010.

[BL07a] C. Bishop and J. Lasserre. “Generative or Discriminative? getting the best of both worlds”. In: *Bayesian Statistics 8*. 2007.

[BL07b] J. A. Bullinaria and J. P. Levy. “Extracting semantic representations from word co-occurrence statistics: a computational study”. en. In: *Behav. Res. Methods* 39.3 (2007), pp. 510–526.

[BL12] J. A. Bullinaria and J. P. Levy. “Extracting semantic representations from word co-occurrence statistics: stop-lists, stem-

ming, and SVD". en. In: *Behav. Res. Methods* 44.3 (2012), pp. 890–907.

[BL88] D. S. Broomhead and D Lowe. "Multivariable Functional Interpolation and Adaptive Networks". In: *Complex Systems* (1988).

[BLK17] O. Bachem, M. Lucic, and A. Krause. "Distributed and provably good seedings for k-means in constant rounds". In: *ICML*. 2017, pp. 292–300.

[Blo20] M. Blondel. *Automatic differentiation*. 2020.

[BLV19] X. Bouthillier, C. Laurent, and P. Vincent. "Unreproducible Research is Reproducible". In: *ICML*. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 725–734.

[BM98] A. Blum and T. Mitchell. "Combining labeled and unlabeled data with co-training". In: *Proceedings of the eleventh annual conference on Computational learning theory*. 1998, pp. 92–100.

[BN01] M. Belkin and P. Niyogi. "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering". In: *NIPS*. 2001, pp. 585–591.

[BNJ03] D. Blei, A. Ng, and M. Jordan. "Latent Dirichlet allocation". In: *JMLR* 3 (2003), pp. 993–1022.

[Bo+08] L. Bo, C. Sminchisescu, A. Kanaujia, and D. Metaxas. "Fast Algorithms for Large Scale Conditional 3D Prediction". In: *CVPR*. 2008.

[Boh92] D. Bohning. "Multinomial logistic regression algorithm". In: *Annals of the Inst. of Statistical Math.* 44 (1992), pp. 197–200.

[Bon13] S. Bonnabel. "Stochastic gradient descent on Riemannian manifolds". In: *IEEE Transactions on Automatic Control* 58.9 (2013), pp. 2217–2229.

[Bos+16] D. Boscaini, J. Masci, E. Rodolà, and M. Bronstein. "Learning shape correspondence with anisotropic convolutional neural networks". In: *Advances in Neural Information Processing Systems*. 2016, pp. 3189–3197.

[Bot+13] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. "Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising". In: *JMLR* 14 (2013), pp. 3207–3260.

[Bow+15] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. "A large annotated corpus for learning natural language inference". In: *EMNLP*. Association for Computational Linguistics, 2015, pp. 632–642.

[BPC20] I. Beltagy, M. E. Peters, and A. Cohan. "Longformer: The Long-Document Transformer". In: *CoRR* abs/2004.05150 (2020). arXiv: 2004.05150.

[Bre01] L. Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

[Bre96] L. Breiman. "Bagging predictors". In: *Machine Learning* 24 (1996), pp. 123–140.

[Bri50] G. W. Brier. "Verification of forecasts expressed in terms of probability". In: *Monthly Weather Review* 78.1 (1950), pp. 1–3.

[Bri90] J. Bridle. "Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition". In: *Neurocomputing: Algorithms, Architectures and Applications*. Ed. by F. F. Soulie and J. Herault. Springer Verlag, 1990, pp. 227–236.

[Bro+17a] M. M. Bronstein, J Bruna, Y LeCun, A Szlam, and P Vandergheynst. "Geometric Deep Learning: Going beyond Euclidean data". In: *IEEE Signal Process. Mag.* 34.4 (2017), pp. 18–42.

[Bro+17b] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. "Geometric deep learning: going beyond euclidean data". In: *IEEE Signal Processing Magazine* 34.4 (2017), pp. 18–42.

[Bro19] J. Brownlee. *Deep Learning for Computer Vision - Machine Learning Mastery*. Accessed: 2020-6-30. Machine Learning Mastery, 2019.

[Bro+20] T. B. Brown et al. "Language Models are Few-Shot Learners". In: (2020). arXiv: 2005.14165 [cs.CL].

[Bro+21] A. Brock, S. De, S. L. Smith, and K. Simonyan. "High-Performance Large-Scale Image Recognition Without Normalization". In: (2021). arXiv: 2102 . 06171 [cs.CV].

[Bro24] W. Brown. *Generative AI Handbook: A Roadmap for Learning Resources*. 2024.

[BRR18] T. D. Bui, S. Ravi, and V. Ramavajjala. "Neural Graph Machines: Learning Neural Networks Using Graphs". In: *WSDM*. 2018.

[Bru+14] J. Bruna, W. Zaremba, A. Szlam, and Y. Lecun. "Spectral networks and locally connected networks on graphs International Conference on Learning Representations (ICLR2014)". In: *CBLS, April* (2014).

[Bru+19] G. Brunner, Y. Liu, D. Pascual, O. Richter, and R. Wattenhofer. "On the Validity of Self-Attention as Explanation in Transformer Models". In: (2019). arXiv: 1908.04211 [cs.CL].

[BS02] M. Balasubramanian and E. L. Schwartz. “The isomap algorithm and topological stability”. en. In: *Science* 295.5552 (2002), p. 7.

[BS16] P. Baldi and P. Sadowski. “A Theory of Local Learning, the Learning Channel, and the Optimality of Backpropagation”. In: *Neural Netw.* 83 (2016), pp. 51–74.

[BS17] D. M. Blei and P. Smyth. “Science and data science”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* (2017).

[BS21] S. Bubeck and M. Sellke. “A Universal Law of Robustness via Isoperimetry”. In: *NIPS* 34 (Dec. 2021), pp. 28811–28822.

[BS94] J. Bernardo and A. Smith. *Bayesian Theory*. John Wiley, 1994.

[BS97] A. J. Bell and T. J. Sejnowski. “The “independent components” of natural scenes are edge filters”. en. In: *Vision Res.* 37.23 (1997), pp. 3327–3338.

[BT04] G. Bouchard and B. Triggs. “The trade-off between generative and discriminative classifiers”. In: *IASC International Symposium on Computational Statistics (COMPSTAT'04)*. 2004, pp. 721–728.

[BT08] D. Bertsekas and J. Tsitsiklis. *Introduction to Probability*. 2nd Edition. Athena Scientific, 2008.

[BT09] A Beck and M Teboulle. “A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems”. In: *SIAM J. Imaging Sci.* 2.1 (2009), pp. 183–202.

[BT73] G. Box and G. Tiao. *Bayesian inference in statistical analysis*. Addison-Wesley, 1973.

[Bul11] A. D. Bull. “Convergence rates of efficient global optimization algorithms”. In: *JMLR* 12 (2011), 2879–2904.

[Bur10] C. J. C. Burges. “Dimension Reduction: A Guided Tour”. en. In: *Foundations and Trends in Machine Learning* (2010).

[Bur25] A. Burkov. *The Hundred-Page Language Models Book*. 2025.

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge, 2004.

[BW08] P. L. Bartlett and M. H. Wegkamp. “Classification with a Reject Option using a Hinge Loss”. In: *JMLR* 9.Aug (2008), pp. 1823–1840.

[BW88] J. Berger and R. Wolpert. *The Likelihood Principle*. 2nd edition. The Institute of Mathematical Statistics, 1988.

[BWL19] Y. Bai, Y.-X. Wang, and E. Liberty. “ProxQuant: Quantized Neural Networks via Proximal Operators”. In: *ICLR*. 2019.

[BY03] P. Buhlmann and B. Yu. “Boosting with the L2 loss: Regression and classification”. In: *JASA* 98.462 (2003), pp. 324–339.

[Byr+16] R Byrd, S Hansen, J Nocedal, and Y Singer. “A Stochastic Quasi-Newton Method for Large-Scale Optimization”. In: *SIAM J. Optim.* 26.2 (2016), pp. 1008–1031.

[BZ20] A. Barbu and S.-C. Zhu. *Monte Carlo Methods*. en. Springer, 2020.

[Cal20] O. Calin. *Deep Learning Architectures: A Mathematical Approach*. en. 1st ed. Springer, 2020.

[Cao+18] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”. In: (2018). arXiv: 1812.08008 [cs.CV].

[CAS16] P. Covington, J. Adams, and E. Sargin. “Deep Neural Networks for YouTube Recommendations”. In: *Proceedings of the 10th ACM Conference on Recommender Systems*. RecSys '16. Association for Computing Machinery, 2016, pp. 191–198.

[CB02] G. Casella and R. Berger. *Statistical inference*. 2nd edition. Duxbury, 2002.

[CBD15] M. Courbariaux, Y. Bengio, and J.-P. David. “BinaryConnect: Training Deep Neural Networks with binary weights during propagations”. In: *NIPS*. 2015.

[CC07] H. Choi and S. Choi. “Robust kernel Isomap”. In: *Pattern Recognit.* 40.3 (2007), pp. 853–862.

[CCD17] B. P. Chamberlain, J. Clough, and M. P. Deisenroth. “Neural embeddings of graphs in hyperbolic space”. In: *arXiv preprint arXiv:1705.10359* (2017).

[CD14] K. Chaudhuri and S. Dasgupta. “Rates of Convergence for Nearest Neighbor Classification”. In: *NIPS*. 2014.

[CD88] W. Cleveland and S. Devlin. “Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting”. In: *JASA* 83.403 (1988), pp. 596–610.

[CDL16] J. Cheng, L. Dong, and M. Lapata. “Long Short-Term Memory-Networks for Machine Reading”. In: *EMNLP*. Association for Computational Linguistics, 2016, pp. 551–561.

[CDL19] S. Chen, E. Dobriban, and J. H. Lee. “Invariance reduces Variance: Understanding Data Augmentation in Deep Learning and Beyond”. In: (2019). arXiv: 1907.10905 [stat.ML].

[CDS02] M. Collins, S. Dasgupta, and R. E. Schapire. “A Generalization of Principal Components Analysis to the Exponential Family”. In: *NIPS-14*. 2002.

[CEL19] Z. Chen, J. B. Estrach, and L. Li. “Supervised community detection with line graph neural networks”. In: *7th Inter-*

national Conference on Learning Representations, ICLR 2019. 2019.

[Cer+17] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation". In: *Proc. 11th Intl. Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, 2017, pp. 1–14.

[CFD10] Y. Cui, X. Z. Fern, and J. G. Dy. "Learning Multiple Nonredundant Clusterings". In: *ACM Transactions on Knowledge Discovery from Data* 4.3 (2010).

[CG16] T. Chen and C. Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *KDD*. ACM, 2016, pp. 785–794.

[CG18] J. Chen and Q. Gu. "Closing the Generalization Gap of Adaptive Gradient Methods in Training Deep Neural Networks". In: (2018). arXiv: 1806.06763 [cs.LG].

[CGG17] S. E. Chazan, J. Goldberger, and S. Gannot. "Speech Enhancement using a Deep Mixture of Experts". In: (2017). arXiv: 1703.09302 [cs.SD].

[CGW21] W. Chen, X. Gong, and Z. Wang. "Neural Architecture Search on ImageNet in Four GPU Hours: A Theoretically Inspired Perspective". In: *ICLR*. 2021.

[CH67] T. Cover and P. Hart. " Nearest neighbor pattern classification". In: *IEEE Trans. Inform. Theory* 13.1 (1967), pp. 21–27.

[CH90] K. W. Church and P. Hanks. "Word Association Norms, Mutual Information, and Lexicography". In: *Computational Linguistics* (1990).

[Cha+17] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina. "Entropy-SGD: Biasing Gradient Descent Into Wide Valleys". In: *ICLR*. 2017.

[Cha+19a] I. Chami, Z. Ying, C. Ré, and J. Leskovec. "Hyperbolic graph convolutional neural networks". In: *Advances in Neural Information Processing Systems*. 2019, pp. 4869–4880.

[Cha+19b] J. J. Chandler, I. Martinez, M. M. Finucane, J. G. Terziev, and A. M. Resch. "Speaking on Data's Behalf: What Researchers Say and How Audiences Choose". en. In: *Eval. Rev.* (2019), p. 193841X19834968.

[Cha+21] I. Chami, S. Abu-El-Haija, B. Perozzi, C. Ré, and K. Murphy. "Machine Learning on Graphs: A Model and Comprehensive Taxonomy". In: *JMLR* (2021).

[Che+16] H.-T. Cheng et al. "Wide & Deep Learning for Recommender Systems". In: (2016). arXiv: 1606.07792 [cs.LG].

[Che+20a] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. "A Simple Framework for Contrastive Learning of Visual Representations". In: *ICML*. 2020.

[Che+20b] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. "A simple framework for contrastive learning of visual representations". In: *ICML*. 2020.

[Che+20c] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. "Big Self-Supervised Models are Strong Semi-Supervised Learners". In: *NIPS*. 2020.

[Chi+19a] W.-L. Chiang, X. Liu, S. Si, Y. Li, S. Bengio, and C.-J. Hsieh. "Cluster-GCN: An Efficient Algorithm for Training Deep and Large Graph Convolutional Networks". In: *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 2019.

[Chi+19b] R. Child, S. Gray, A. Radford, and I. Sutskever. "Generating Long Sequences with Sparse Transformers". In: *CoRR* abs/1904.10509 (2019). arXiv: 1904 . 10509.

[CHL05] S. Chopra, R. Hadsell, and Y. LeCun. "Learning a Similarity Metric Discriminatively, with Application to Face Verification". en. In: *CVPR*. 2005.

[Cho+14a] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". In: *EMNLP*. 2014.

[Cho+14b] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. "On the properties of neural machine translation: Encoder-decoder approaches". In: *arXiv preprint arXiv:1409.1259* (2014).

[Cho+15] Y. Chow, A. Tamar, S. Mannor, and M. Pavone. "Risk-Sensitive and Robust Decision-Making: a CVaR Optimization Approach". In: *NIPS*. 2015, pp. 1522–1530.

[Cho17] F. Chollet. *Deep learning with Python*. Manning, 2017.

[Cho+19] K. Choromanski, M. Rowland, W. Chen, and A. Weller. "Unifying Orthogonal Monte Carlo Methods". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 1203–1212.

[Cho+20a] K. Choromanski et al. "Masked Language Modeling for Proteins via Linearly Scalable Long-Context Transform-

ers". In: (2020). arXiv: 2006 . 03555 [cs.LG].

[Cho+20b] K. Choromanski et al. "Rethinking Attention with Performers". In: *CoRR* abs/2009.14794 (2020). arXiv: 2009 . 14794.

[Cho21] F. Chollet. *Deep learning with Python (second edition)*. Manning, 2021.

[Chu+15] J. Chung, K. Kastner, L. Dinh, K. Goel, A. Courville, and Y. Bengio. "A Recurrent Latent Variable Model for Sequential Data". In: *NIPS*. 2015.

[Chu+22] H. W. Chung et al. "Scaling Instruction-Finetuned Language Models". In: (Oct. 2022). arXiv: 2210.11416 [cs.LG].

[Chu97] F. Chung. *Spectral Graph Theory*. AMS, 1997.

[Cir+10] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber. "Deep Big Simple Neural Nets For Handwritten Digit Recognition". In: *Neural Computation* 22.12 (2010), pp. 3207–3220.

[Cir+11] D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber. "Flexible, High Performance Convolutional Neural Networks for Image Classification". In: *IJCAI*. 2011.

[CL96] B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, 1996.

[Cla21] A. Clayton. *Bernoulli's Fallacy: Statistical Illogic and the Crisis of Modern Science*. en. Columbia University Press, 2021.

[CLX15] S. Cao, W. Lu, and Q. Xu. "Grarep: Learning graph representations with global structural information". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM. 2015, pp. 891–900.

[CNB17] C. Chelba, M. Norouzi, and S. Bengio. "N-gram Language Modeling using Recurrent Neural Network Estimation". In: (2017). arXiv: 1703.10724 [cs.CL].

[Coh+17] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik. "EMNIST: an extension of MNIST to handwritten letters". In: (2017). arXiv: 1702.05373 [cs.CV].

[Coh94] J. Cohen. "The earth is round ($p < .05$)". In: *American Psychologist* 49.12 (1994), pp. 997–1003.

[Con+17] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes. "Supervised learning of universal sentence representations from natural language inference data". In: *arXiv preprint arXiv:1705.02364* (2017).

[Coo05] J. Cook. *Exact Calculation of Beta Inequalities*. Tech. rep. M. D. Anderson Cancer Center, Dept. Biostatistics, 2005.

[Cor+16] C. Cortes, X. Gonzalvo, V. Kuznetsov, M. Mohri, and S. Yang. "AdaNet: Adaptive Structural Learning of Artificial Neural Networks". In: (2016). arXiv: 1607 . 01097 [cs.LG].

[CP10] M. A. Carreira-Perpinan. "The Elastic Embedding Algorithm for Dimensionality Reduction". In: *ICML*. 2010.

[CP19] A. Coenen and A. Pearce. *Understanding UMAP*. 2019.

[CPS06] K. Chellapilla, S. Puri, and P. Simard. "High Performance Convolutional Neural Networks for Document Processing". In: *10th Intl. Workshop on Frontiers in Handwriting Recognition*. 2006.

[CRW17] K. Choromanski, M. Rowland, and A. Weller. "The Unreasonable Effectiveness of Structured Random Orthogonal Embeddings". In: *NIPS*. 2017.

[CS20] F. E. Curtis and K Scheinberg. "Adaptive Stochastic Optimization: A Framework for Analyzing Stochastic Optimization Algorithms". In: *IEEE Signal Process. Mag.* 37.5 (2020), pp. 32–42.

[Csu17] G. Csurka. "Domain Adaptation for Visual Applications: A Comprehensive Survey". In: *Domain Adaptation in Computer Vision Applications*. Ed. by G. Csurka. 2017.

[CT06] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. 2nd edition. John Wiley, 2006.

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley, 1991.

[Cub+19] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. "AutoAugment: Learning Augmentation Policies from Data". In: *CVPR*. 2019.

[CUH16] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)". In: *ICLR*. 2016.

[Cui+19] X. Cui, K. Zheng, L. Gao, B. Zhang, D. Yang, and J. Ren. "Multiscale Spatial-Spectral Convolutional Network with Image-Based Framework for Hyperspectral Imagery Classification". en. In: *Remote Sensing* 11.19 (2019), p. 2220.

[Cur+17] J. D. Curtó, I. C. Zarza, F Yang, A Smola, F Torre, C. W. Ngo, and L Gool. "McKernel: A Library for Approximate Kernel Expansions in Log-linear Time". In: (2017). arXiv: 1702.08159v14 [cs.LG].

[Cyb89] G. Cybenko. "Approximation by superpositions of a sigmoidal function". In:

Mathematics of Control, Signals, and Systems 2 (1989), 303–331.

[D'A+20] A. D'Amour et al. "Underspecification Presents Challenges for Credibility in Modern Machine Learning". In: (2020). arXiv: 2011.03395 [cs.LG].

[Dah+11] G. E. Dahl, D. Yu, L. Deng, and A. Acero. "Large vocabulary continuous speech recognition with context-dependent DBN-HMMS". In: *ICASSP*. IEEE, 2011, pp. 4688–4691.

[Dai+19] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov. "Transformer-XL: Attentive Language Models beyond a Fixed-Length Context". In: *Proc. ACL*. 2019, pp. 2978–2988.

[Dao+19] T. Dao, A. Gu, A. J. Ratner, V. Smith, C. De Sa, and C. Re. "A Kernel Theory of Modern Data Augmentation". In: *ICML*. 2019.

[Dau17] J. Daunizeau. "Semi-analytical approximations to statistical moments of sigmoid and softmax mappings of normal variables". In: (2017). arXiv: 1703.00091 [stat.ML].

[Day+95] P. Dayan, G. Hinton, R. Neal, and R. Zemel. "The Helmholtz machine". In: *Neural Networks* 9.8 (1995).

[DB18] A. Defazio and L. Bottou. "On the Ineffectiveness of Variance Reduced Optimization for Deep Learning". In: (2018). arXiv: 1812.04529 [cs.LG].

[DBLJ14] A. Defazio, F. Bach, and S. Lacoste-Julien. "SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives". In: *NIPS*. Curran Associates, Inc., 2014, pp. 1646–1654.

[DDDM04] I Daubechies, M Defrise, and C De Mol. "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint". In: *Commun. Pure Appl. Math.* Advances in E 57.11 (2004), pp. 1413–1457.

[Dee+90] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. "Indexing by Latent Semantic Analysis". In: *J. of the American Society for Information Science* 41.6 (1990), pp. 391–407.

[DeG70] M. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, 1970.

[Den+12] J. Deng, J Krause, A. C. Berg, and L. Fei-Fei. "Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition". In: *CVPR*. 2012, pp. 3450–3457.

[Den+14] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam. "Large-Scale Object Classification using Label Relation Graphs". In: *ECCV*. 2014.

[Dev+19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL*. 2019.

[DG06] J. Davis and M. Goadrich. "The Relationship Between Precision-Recall and ROC Curves". In: *ICML*. 2006, pp. 233–240.

[DHS01] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. 2nd edition. Wiley Interscience, 2001.

[DHS11] J. Duchi, E. Hazan, and Y. Singer. "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". In: *JMLR* 12 (2011), pp. 2121–2159.

[Die98] T. G. Dietterich. "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms". In: *Neural Computation*. 10.7 (1998), pp. 1895–1923.

[Din+15] N. Ding, J. Deng, K. Murphy, and H. Neven. "Probabilistic Label Relation Graphs with Ising Models". In: *ICCV*. 2015.

[DJ15] S. Dray and J. Josse. "Principal component analysis with missing values: a comparative survey of methods". In: *Plant Ecol.* 216.5 (2015), pp. 657–667.

[DKK12] G Dror, N Koenigstein, and Y Koren. "Web-Scale Media Recommendation Systems". In: *Proc. IEEE* 100.9 (2012), pp. 2722–2736.

[DLLP97] T. Dietterich, R. Lathrop, and T. Lozano-Perez. "Solving the multiple instance problem with axis-parallel rectangles". In: *Artificial Intelligence* 89 (1997), pp. 31–71.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *J. of the Royal Statistical Society, Series B* 34 (1977), pp. 1–38.

[DM01] D. van Dyk and X.-L. Meng. "The Art of Data Augmentation". In: *J. Computational and Graphical Statistics* 10.1 (2001), pp. 1–50.

[DM16] P. Drineas and M. W. Mahoney. "RandNLA: Randomized Numerical Linear Algebra". In: *CACM* (2016).

[DMB21] Y. Dar, V. Muthukumar, and R. G. Baraniuk. "A Farewell to the Bias-Variance Tradeoff? An Overview of the Theory of Overparameterized Machine Learning". In: (Sept. 2021). arXiv: 2109.02355 [stat.ML].

[Do+19] T.-T. Do, T. Tran, I. Reid, V. Kumar, T. Hoang, and G. Carneiro. "A Theoret-

ically Sound Upper Bound on the Triplet Loss for Improving the Efficiency of Deep Distance Metric Learning". In: *CVPR*. 2019, pp. 10404–10413.

[Doe16] C. Doersch. "Tutorial on Variational Autoencoders". In: (2016). arXiv: 1606 . 05908 [stat.ML].

[Don95] D. L. Donoho. "De-noising by soft-thresholding". In: *IEEE Trans. Inf. Theory* 41.3 (1995), pp. 613–627.

[Dos+21] A. Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *ICLR*. 2021.

[Doy+07] K. Doya, S. Ishii, A. Pouget, and R. P. N. Rao, eds. *Bayesian Brain: Probabilistic Approaches to Neural Coding*. MIT Press, 2007.

[DP97] P. Domingos and M. Pazzani. "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss". In: *Machine Learning* 29 (1997), pp. 103–130.

[DR21] H. Duanmu and D. M. Roy. "On extended admissibale procedures and their nonstandard Bayes risk". In: *Annals of Statistics* (2021).

[Dri+04] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. "Clustering Large Graphs via the Singular Value Decomposition". In: *Machine Learning* 56 (2004), pp. 9–33.

[DS12] M. Der and L. K. Saul. "Latent Coincidence Analysis: A Hidden Variable Model for Distance Metric Learning". In: *NIPS*. Curran Associates, Inc., 2012, pp. 3230–3238.

[DSK16] V. Dumoulin, J. Shlens, and M. Kudlur. "A Learned Representation For Artistic Style". In: (2016). arXiv: 1610 . 07629 [cs.CV].

[Dum+18] A. Dumitrache, O. Inel, B. Timmermans, C. Ortiz, R.-J. Sips, L. Aroyo, and C. Welty. "Empirical Methodology for Crowdsourcing Ground Truth". In: *Semantic Web Journal* (2018).

[Duv14] D. Duvenaud. "Automatic Model Construction with Gaussian Processes". PhD thesis. Computational and Biological Learning Laboratory, University of Cambridge, 2014.

[DV16] V. Dumoulin and F. Visin. "A guide to convolution arithmetic for deep learning". In: (2016). arXiv: 1603 . 07285 [stat.ML].

[Dwi+23] R. Dwivedi, C. Singh, B. Yu, and M. J. Wainwright. "Revisiting minimum description length complexity in overparameterized models". In: *J. Mach. Learn. Res.* (2023).

[Dzi+24] N. Dziri et al. "Faith and Fate: Limits of Transformers on Compositionality". In: *NIPS*. 2024.

[EDH19] K. Ethayarajh, D. Duvenaud, and G. Hirst. "Towards Understanding Linear Word Analogies". In: *Proc. ACL*. Association for Computational Linguistics, 2019, pp. 3253–3262.

[EF15] D. Eigen and R. Fergus. "Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture". In: *ICCV*. 2015.

[Efr+04] B. Efron, I. Johnstone, T. Hastie, and R. Tibshirani. "Least angle regression". In: *Annals of Statistics* 32.2 (2004), pp. 407–499.

[Efr86] B. Efron. "Why Isn't Everyone a Bayesian?" In: *The American Statistician* 40.1 (1986).

[Ein16] A Einstein. "Die Grundlage der allgemeinen Relativitätstheorie". In: *Ann. Phys.* 354.7 (1916), pp. 769–822.

[Eis19] J. Eisenstein. *Introduction to Natural Language Processing*. 2019.

[Elk03] C. Elkan. "Using the triangle inequality to accelerate k-means". In: *ICML*. 2003.

[EMH19] T. Elsken, J. H. Metzen, and F. Hutter. "Neural Architecture Search: A Survey". In: *JMLR* 20 (2019), pp. 1–21.

[Erh+10] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio. "Why Does Unsupervised Pre-training Help Deep Learning?" In: *JMLR* 11 (2010), pp. 625–660.

[FAL17] C. Finn, P. Abbeel, and S. Levine. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In: *ICML*. 2017.

[Fen+21] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy. "A Survey of Data Augmentation Approaches for NLP". In: (2021). arXiv: 2105.03075 [cs.CL].

[FH20] E. Fong and C. Holmes. "On the marginal likelihood and cross-validation". In: *Biometrika* 107.2 (2020).

[FHK12] A. Feuerverger, Y. He, and S. Khatri. "Statistical Significance of the Netflix Challenge". In: *Stat. Sci.* 27.2 (2012), pp. 202–231.

[FHT00] J. Friedman, T. Hastie, and R. Tibshirani. "Additive logistic regression: a statistical view of boosting". In: *Annals of statistics* 28.2 (2000), pp. 337–374.

[FHT10] J. Friedman, T. Hastie, and R. Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *J. of Statistical Software* 33.1 (2010).

[Fir57] J. Firth. “A synopsis of linguistic theory 1930-1955”. In: *Studies in Linguistic Analysis*. Ed. by F. Palmer. 1957.

[FJ02] M. A. T. Figueiredo and A. K. Jain. “Unsupervised Learning of Finite Mixture Models”. In: *IEEE PAMI* 24.3 (2002), pp. 381–396.

[FM03] J. H. Friedman and J. J. Meulman. “Multiple additive regression trees with application in epidemiology”. en. In: *Stat. Med.* 22.9 (2003), pp. 1365–1381.

[FMN16] C. Fefferman, S. Mitter, and H. Narayanan. “Testing the manifold hypothesis”. In: *J. Amer. Math. Soc.* 29.4 (2016), pp. 983–1049.

[FNW07] M. Figueiredo, R. Nowak, and S. Wright. “Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems”. In: *IEEE. J. on Selected Topics in Signal Processing* (2007).

[For+21] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. “Sharpness-aware Minimization for Efficiently Improving Generalization”. In: *ICLR*. 2021.

[Fos19] D. Foster. *Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play*. 1 edition. O’Reilly Media, 2019.

[FR07] C. Fraley and A. Raftery. “Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering”. In: *J. of Classification* 24 (2007), pp. 155–181.

[Fra+17] L. Franceschi, M. Donini, P. Frasconi, and M. Pontil. “Forward and Reverse Gradient-Based Hyperparameter Optimization”. In: *ICML*. 2017.

[Fre98] B. Frey. *Graphical Models for Machine Learning and Digital Communication*. MIT Press, 1998.

[Fri01] J. Friedman. “Greedy Function Approximation: a Gradient Boosting Machine”. In: *Annals of Statistics* 29 (2001), pp. 1189–1232.

[Fri97a] J. Friedman. “On bias, variance, 0-1 loss and the curse of dimensionality”. In: *J. Data Mining and Knowledge Discovery* 1 (1997), pp. 55–77.

[Fri97b] J. H. Friedman. “Data mining and statistics: What’s the connection”. In: *Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics*. 1997.

[Fri99] J. Friedman. *Stochastic Gradient Boosting*. Tech. rep. 1999.

[FS96] Y. Freund and R. R. Schapire. “Experiments with a new boosting algorithm”. In: *ICML*. 1996.

[FT05] M. Fashing and C. Tomasi. “Mean shift is a bound optimization”. en. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 27.3 (2005), pp. 471–474.

[Fu98] W. Fu. “Penalized regressions: the bridge versus the lasso”. In: *J. Computational and graphical statistics* 7 (1998), 397–416.

[Fuk75] K. Fukushima. “Cognitron: a self-organizing multilayered neural network”. In: *Biological Cybernetics* 20.6 (1975), pp. 121–136.

[Fuk80] K Fukushima. “Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position”. en. In: *Biol. Cybern.* 36.4 (1980), pp. 193–202.

[Fuk90] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. 2nd edition. Academic Press, 1990.

[Gag94] P. Gage. “A New Algorithm for Data Compression”. In: *Dr Dobbs Journal* (1994).

[Gan+16] Y Ganin, E Ustinova, H Ajakan, P Germain, and others. “Domain-adversarial training of neural networks”. In: *JMLR* (2016).

[Gär03] T. Gärtner. “A Survey of Kernels for Structured Data”. In: *SIGKDD Explor. Newsl.* 5.1 (2003), pp. 49–58.

[Gar+18] J. Gardner, G. Pleiss, K. Q. Weinberger, D. Bindel, and A. G. Wilson. “GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration”. In: *NIPS*. Ed. by S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett. Curran Associates, Inc., 2018, pp. 7576–7586.

[GASG18] D. G. A. Smith and J. Gray. “opt_einsum - A Python package for optimizing contraction order for einsum-like expressions”. In: *JOSS* 3.26 (2018), p. 753.

[GB05] Y. Grandvalet and Y. Bengio. “Semi-supervised learning by entropy minimization”. In: *Advances in neural information processing systems*. 2005, pp. 529–536.

[GB10] X. Glorot and Y. Bengio. “Understanding the difficulty of training deep feedforward neural networks”. In: *AISTATS*. 2010, pp. 249–256.

[GB18] V. Garcia and J. Bruna. “Few-shot Learning with Graph Neural Networks”. In: *International Conference on Learning Representations (ICLR)*. 2018.

[GBB11] X. Glorot, A. Bordes, and Y. Bengio. “Deep Sparse Rectifer Neural Networks”. In: *AISTATS*. 2011.

[GBC16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[GBD92] S. Geman, E. Bienenstock, and R. Doursat. “Neural networks and the bias-variance dilemma”. In: *Neural Computing* 4 (1992), pp. 1–58.

[GC20] A. Gelman and B. Carpenter. “Bayesian analysis of tests with unknown specificity and sensitivity”. In: *J. of Royal Stat. Soc. Series C* medrxiv;2020.05.22.20108944v2 (2020).

[GEB16] L. A. Gatys, A. S. Ecker, and M. Bethge. “Image style transfer using convolutional neural networks”. In: *CVPR*. 2016, pp. 2414–2423.

[GEH19] T. Gale, E. Elsen, and S. Hooker. “The State of Sparsity in Deep Neural Networks”. In: (2019). arXiv: 1902.09574 [cs.LG].

[Gel+04] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian data analysis*. 2nd edition. Chapman and Hall, 2004.

[Gel+14] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis, Third Edition*. Third edition. Chapman and Hall/CRC, 2014.

[Gel16] A. Gelman. “The problems with p-values are not just with p-values”. In: *American Statistician* (2016).

[Gér17] A. Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques for Building Intelligent Systems*. en. O’Reilly Media, Incorporated, 2017.

[Gér19] A. Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques for Building Intelligent Systems (2nd edition)*. en. O’Reilly Media, Incorporated, 2019.

[GEY19] Y. Geifman and R. El-Yaniv. “SelectiveNet: A Deep Neural Network with an Integrated Reject Option”. In: *ICML*. 2019.

[GG16] Y. Gal and Z. Ghahramani. “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning”. In: *ICML*. 2016.

[GH96] Z. Ghahramani and G. Hinton. *The EM Algorithm for Mixtures of Factor Analyzers*. Tech. rep. Dept. of Comp. Sci., Uni. Toronto, 1996.

[GHV14] A. Gelman, J. Hwang, and A. Vehtari. “Understanding predictive information criteria for Bayesian models”. In: *Statistics and Computing* 24.6 (2014), pp. 997–1016.

[Gib97] M. Gibbs. “Bayesian Gaussian Processes for Regression and Classification”. PhD thesis. U. Cambridge, 1997.

[Gil+17] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. “Neural message passing for quantum chemistry”. In: *ICML*. 2017, pp. 1263–1272.

[Gil+21] J. Gilmer, B. Ghorbani, A. Garg, S. Kudugunta, B. Neyshabur, D. Cardoze, G. Dahl, Z. Nado, and O. Firat. “A Loss Curvature Perspective on Training Instability in Deep Learning”. In: (2021). arXiv: 2110.04369 [cs.LG].

[GIM99] A. Gionis, P. Indyk, and R. Motwani. “Similarity Search in High Dimensions via Hashing”. In: *Proc. 25th Intl. Conf. on Very Large Data Bases*. VLDB ’99. 1999, pp. 518–529.

[GKS18] V. Gupta, T. Koren, and Y. Singer. “Shampoo: Preconditioned Stochastic Tensor Optimization”. In: *ICML*. 2018.

[GL15] B. Gu and C. Ling. “A New Generalized Error Path Algorithm for Model Selection”. In: *ICML*. 2015.

[GL16] A. Grover and J. Leskovec. “node2vec: Scalable feature learning for networks”. In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2016, pp. 855–864.

[GMS05] M. Gori, G. Monfardini, and F. Scarselli. “A new model for learning in graph domains”. In: *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005*. Vol. 2. IEEE. 2005, pp. 729–734.

[GNK18] R. A. Güler, N. Neverova, and I. Kokkinos. “Densepose: Dense human pose estimation in the wild”. In: *CVPR*. 2018, pp. 7297–7306.

[God18] P. Godec. *Graph Embeddings; The Summary*. https://towardsdatascience.com/graph-embeddings-the-summary-cc6075aba007. 2018.

[GOF18] O. Gouvert, T. Oberlin, and C. Févotte. “Negative Binomial Matrix Factorization for Recommender Systems”. In: (2018). arXiv: 1801.01708 [cs.LG].

[Gol+01] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. “Eigentaste: A Constant Time Collaborative Filtering Algorithm”. In: *Information Retrieval* 4.2 (2001), pp. 133–151.

[Gol+05] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. “Neighbourhood Components Analysis”. In: *NIPS*. 2005.

[Gol+92] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. “Using collaborative filtering to weave an information tapestry”. In: *Commun. ACM* 35.12 (1992), pp. 61–70.

[Gon85] T. Gonzales. “Clustering to minimize the maximum intercluster distance”. In:

Theor. Comp. Sci. 38 (1985), pp. 293–306.

[Goo01] N. Goodman. "Classes for fast maximum entropy training". In: *ICASSP*. 2001.

[Goo+14] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative Adversarial Networks". In: *NIPS*. 2014.

[Gor06] P. F. Gorder. "Neural Networks Show New Promise for Machine Vision". In: *Computing in science & engineering* 8.6 (2006), pp. 4–8.

[Got+19] A. Gotmare, N. S. Keskar, C. Xiong, and R. Socher. "A Closer Look at Deep Learning Heuristics: Learning rate restarts, Warmup and Distillation". In: *ICLR*. 2019.

[GOV18] W Gao, S Oh, and P Viswanath. "Demystifying Fixed k -Nearest Neighbor Information Estimators". In: *IEEE Trans. Inf. Theory* 64.8 (2018), pp. 5629–5661.

[GR07] T. Gneiting and A. E. Raftery. "Strictly Proper Scoring Rules, Prediction, and Estimation". In: *JASA* 102.477 (2007), pp. 359–378.

[GR18] A. Graves and M.-A. Ranzato. "Tutorial on unsupervised deep learning: part 2". In: *NIPS*. 2018.

[GR19] P. Grünwald and T. Roos. "Minimum description length revisited". In: *Int. J. Math. Ind.* 11.01 (Dec. 2019), p. 1930001.

[Gra04] Y. Grandvalet. "Bagging Equalizes Influence". In: *Mach. Learn.* 55 (2004), pp. 251–270.

[Gra11] A. Graves. "Practical variational inference for neural networks". In: *Advances in neural information processing systems*. 2011, pp. 2348–2356.

[Gra13] A. Graves. "Generating Sequences With Recurrent Neural Networks". In: (2013). arXiv: 1308.0850 [cs.NE].

[Gra+17] E. Grave, A. Joulin, M. Cissé, D. Grangier, and H. Jégou. "Efficient softmax approximation for GPUs". In: *ICML*. 2017.

[Gra+18] E. Grant, C. Finn, S. Levine, T. Darrell, and T. Griffiths. "Recasting Gradient-Based Meta-Learning as Hierarchical Bayes". In: *ICLR*. 2018.

[Gra+20] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky. "Your classifier is secretly an energy based model and you should treat it like one". In: *ICLR*. 2020.

[Gre+17] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber. "LSTM: A Search Space Odyssey". In: *IEEE Transactions on Neural Networks and Learning Systems* 28.10 (2017).

[Gri20] T. L. Griffiths. "Understanding Human Intelligence through Human Limitations". en. In: *Trends Cogn. Sci.* 24.11 (2020), pp. 873–883.

[Gru07] P. Grunwald. *The Minimum Description Length Principle*. MIT Press, 2007.

[GS08] Y Guo and D Schuurmans. "Efficient global optimization for exponential family PCA and low-rank matrix factorization". In: *2008 46th Annual Allerton Conference on Communication, Control, and Computing*. 2008, pp. 1100–1107.

[GS97] C. M. Grinstead and J. L. Snell. *Introduction to probability (2nd edition)*. American Mathematical Society, 1997.

[GSK18] S. Gidaris, P. Singh, and N. Komodakis. "Unsupervised Representation Learning by Predicting Image Rotations". In: *ICLR*. 2018.

[GTA00] G. Gigerenzer, P. M. Todd, and ABC Research Group. *Simple Heuristics That Make Us Smart*. en. Illustrated edition. Oxford University Press, 2000.

[Gu+18] A. Gu, F. Sala, B. Gunel, and C. Ré. "Learning Mixed-Curvature Representations in Product Spaces". In: *International Conference on Learning Representations* (2018).

[Gua+10] Y. Guan, J. Dy, D. Niu, and Z. Ghahramani. "Variational Inference for Nonparametric Multiple Clustering". In: *1st Intl. Workshop on Discovering, Summarizing and Using Multiple Clustering (MultiClust)*. 2010.

[Gua+17] S. Guadarrama, R. Dahl, D. Bieber, M. Norouzi, J. Shlens, and K. Murphy. "PixColor: Pixel Recursive Colorization". In: *BMVC*. 2017.

[Gul+20] A. Gulati et al. "Conformer: Convolution-augmented Transformer for Speech Recognition". In: (2020). arXiv: 2005.08100 [eess.AS].

[Guo09] Y. Guo. "Supervised exponential family principal component analysis via convex optimization". In: *NIPS*. 2009.

[Guo+17] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He. "DeepFM: a factorization-machine based neural network for CTR prediction". In: *IJCAI*. IJCAI'17. AAAI Press, 2017, pp. 1725–1731.

[Gus01] M. Gustafsson. "A probabilistic derivation of the partial least-squares algorithm". In: *Journal of Chemical Information and Modeling* 41 (2001), pp. 288–294.

[GVZ16] A. Gupta, A. Vedaldi, and A. Zisserman. "Synthetic Data for Text Localisation in Natural Images". In: *CVPR*. 2016.

[GZE19] A. Grover, A. Zweig, and S. Ermon. "Graphite: Iterative Generative Modeling of Graphs". In: *International Conference*

on Machine Learning. 2019, pp. 2434–2444.

[HA85] L. Hubert and P. Arabie. "Comparing Partitions". In: *J. of Classification* 2 (1985), pp. 193–218.

[HAB19] M. Hein, M. Andriushchenko, and J. Bitterwolf. "Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem". In: *CVPR*. 2019.

[Hac75] I. Hacking. *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. Cambridge University Press, 1975.

[Háj08] A. Hájek. "Dutch Book Arguments". In: *The Oxford Handbook of Rational and Social Choice*. Ed. by P. Anand, P. Pattanaik, and C. Puppe. Oxford University Press, 2008.

[Han+20] B. Han, Q. Yao, T. Liu, G. Niu, I. W. Tsang, J. T. Kwok, and M. Sugiyama. "A Survey of Label-noise Representation Learning: Past, Present and Future". In: (2020). arXiv: 2011.04406 [cs.LG].

[Has+04] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. "The entire regularization path for the support vector machine". In: *JMLR* 5 (2004), pp. 1391–1415.

[Has+09] T. Hastie, S. Rosset, J. Zhu, and H. Zou. "Multi-class AdaBoost". In: *Statistics and tis Interface* 2.3 (2009), pp. 349–360.

[Has+17] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick. "Neuroscience-Inspired Artificial Intelligence". en. In: *Neuron* 95.2 (2017), pp. 245–258.

[Has87] J. Hastad. *Computational limits of small-depth circuits*. MIT Press, 1987.

[HB17] X. Huang and S. Belongie. "Arbitrary style transfer in real-time with adaptive instance normalization". In: *ICCV*. 2017.

[HBK23] P. Haluptzok, M. Bowers, and A. T. Kalai. "Language Models Can Teach Themselves to Program Better". In: *ICLR*. Feb. 2023.

[HCD12] D. Hoiem, Y. Chodpathumwan, and Q. Dai. "Diagnosing Error in Object Detectors". In: *ECCV*. 2012.

[HCL03] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. *A Practical Guide to Support Vector Classification*. Tech. rep. Dept. Comp. Sci., National Taiwan University, 2003.

[He+15] K. He, X. Zhang, S. Ren, and J. Sun. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *ICCV*. 2015.

[He+16a] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". In: *CVPR*. 2016.

[He+16b] K. He, X. Zhang, S. Ren, and J. Sun. "Identity Mappings in Deep Residual Networks". In: *ECCV*. 2016.

[He+17] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua. "Neural Collaborative Filtering". In: *WWW*. 2017.

[HE18] D. Ha and D. Eck. "A Neural Representation of Sketch Drawings". In: *ICLR*. 2018.

[He+20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. "Momentum contrast for unsupervised visual representation learning". In: *CVPR*. 2020, pp. 9729–9738.

[HG16] D. Hendrycks and K. Gimpel. "Gaussian Error Linear Units (GELUs)". In: *arXiv [cs.LG]* (2016).

[HG20] J. Howard and S. Gugger. *Deep Learning for Coders with Fastai and PyTorch: AI Applications Without a PhD*. en. 1st ed. O'Reilly Media, 2020.

[HGD19] K. He, R. Girshick, and P. Dollár. "Rethinking ImageNet Pre-training". In: *CVPR*. 2019.

[Hin+12] G. E. Hinton et al. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups". In: *IEEE Signal Process. Mag.* 29.6 (2012), pp. 82–97.

[Hin13] G. Hinton. *CSC 2535 Lecture 11: Nonlinear dimensionality reduction*. 2013.

[Hin14] G. Hinton. *Lecture 6e on neural networks (RMSprop: Divide the gradient by a running average of its recent magnitude)*. 2014.

[HK15] F. M. Harper and J. A. Konstan. "The MovieLens Datasets: History and Context". In: *ACM Trans. Interact. Intell. Syst.* 5.4 (2015), pp. 1–19.

[HK92] A Hertz and J Krogh. "Generalization in a linear perceptron in the presence of noise". In: *J. Physics A* (1992).

[HL04] D. R. Hunter and K. Lange. "A Tutorial on MM Algorithms". In: *The American Statistician* 58 (2004), pp. 30–37.

[HMT11] N. Halko, P.-G. Martinsson, and J. A. Tropp. "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions". In: *SIAM Rev., Survey and Review section* 53.2 (2011), pp. 217–288.

[HN19] C. M. Holmes and I. Nemenman. "Estimation of mutual information for real-valued data with error bars and controlled bias". en. In: *Phys Rev E* 100.2-1 (2019), p. 022404.

[Hoc+01] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber. "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies". In: *A Field Guide to Dynamical Recurrent Neural*

Networks. Ed. by S. C. Kremer and J. F. Kolen. 2001.

[Hoe+14] R. Hoekstra, R. D. Morey, J. N. Rouder, and E.-J. Wagenmakers. "Robust misinterpretation of confidence intervals". en. In: *Psychon. Bull. Rev.* 21.5 (2014), pp. 1157–1164.

[Hoe+21] T. Hoefler, D. Alistarh, T. Ben-Nun, N. Dryden, and A. Peste. "Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks". In: (2021). arXiv: 2102.00554 [cs.LG].

[Hof09] P. D. Hoff. *A First Course in Bayesian Statistical Methods*. Springer, 2009.

[Hor61] P Horst. "Generalized canonical correlations and their applications to experimental data". en. In: *J. Clin. Psychol.* 17 (1961), pp. 331–347.

[Hor91] K. Hornik. "Approximation Capabilities of Multilayer Feedforward Networks". In: *Neural Networks* 4.2 (1991), pp. 251–257.

[Hos+19] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga. "A Comprehensive Survey of Deep Learning for Image Captioning". In: *ACM Computing Surveys* (2019).

[HOT06] G. Hinton, S. Osindero, and Y. Teh. "A fast learning algorithm for deep belief nets". In: *Neural Computation* 18 (2006), pp. 1527–1554.

[Hot36] H. Hotelling. "Relations Between Two Sets of Variates". In: *Biometrika* 28.3/4 (1936), pp. 321–377.

[Hou+12] N. Houlsby, F. Huszar, Z. Ghahramani, and J. M. Hernández-lobato. "Collaborative Gaussian Processes for Preference Learning". In: *NIPS*. 2012, pp. 2096–2104.

[Hou+19] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. "Parameter-Efficient Transfer Learning for NLP". In: *ICML*. 2019.

[How+17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications". In: *CVPR*. 2017.

[HR03] G. E. Hinton and S. T. Roweis. "Stochastic Neighbor Embedding". In: *NIPS*. 2003, pp. 857–864.

[HR76] L. Hyafil and R. Rivest. "Constructing Optimal Binary Decision Trees is NP-complete". In: *Information Processing Letters* 5.1 (1976), pp. 15–17.

[HRP21] M. Huisman, J. N. van Rijn, and A. Plaat. "A Survey of Deep Meta-Learning". In: *AI Review* (2021).

[HS19] J. Haochen and S. Sra. "Random Shuffling Beats SGD after Finite Epochs". In: *ICML*. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 2624–2633.

[HS97a] S Hochreiter and J Schmidhuber. "Flat minima". en. In: *Neural Comput.* 9.1 (1997), pp. 1–42.

[HS97b] S. Hochreiter and J. Schmidhuber. "Long short-term memory". In: *Neural Computation* 9.8 (1997), 1735–1780.

[HSW89] K. Hornik, M. Stinchcombe, and H. White. "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5 (1989), pp. 359–366.

[HT90] T. Hastie and R. Tibshirani. *Generalized additive models*. Chapman and Hall, 1990.

[HTF01] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. 2nd edition. Springer, 2009.

[HTW15] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015.

[Hua14] G.-B. Huang. "An Insight into Extreme Learning Machines: Random Neurons, Random Features and Kernels". In: *Cognit. Comput.* 6.3 (2014), pp. 376–390.

[Hua+17a] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. "Densely Connected Convolutional Networks". In: *CVPR*. 2017.

[Hua+17b] J. Huang et al. "Speed/accuracy trade-offs for modern convolutional object detectors". In: *CVPR*. 2017.

[Hua+18] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck. "Music Transformer". In: (2018). arXiv: 1809.04281 [cs.LG].

[Hub+08] M. F. Huber, T Bailey, H Durrant-Whyte, and U. D. Hanebeck. "On entropy approximation for Gaussian mixture random vectors". In: *2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*. 2008, pp. 181–188.

[Hub64] P. Huber. "Robust Estimation of a Location Parameter". In: *Annals of Statistics* 53 (1964), 73–101.

[Hut90] M. F. Hutchinson. "A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines". In: *Communications in Statistics - Simulation*

and Computation 19.2 (1990), pp. 433–450.

[HVD14] G. Hinton, O. Vinyals, and J. Dean. "Distilling the Knowledge in a Neural Network". In: *NIPS Deep Learning Workshop*. 2014.

[HW62] D. Hubel and T. Wiesel. "Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex". In: *J. Physiology* 160 (1962), pp. 106–154.

[HY01] M. Hansen and B. Yu. "Model selection and the principle of minimum description length". In: *JASA* (2001).

[HYL17] W. Hamilton, Z. Ying, and J. Leskovec. "Inductive representation learning on large graphs". In: *Advances in Neural Information Processing Systems*. 2017, pp. 1024–1034.

[Idr+17] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah. "The THUMOS challenge on action recognition for videos "in the wild"". In: *Comput. Vis. Image Underst.* 155 (2017), pp. 1–23.

[Ie+19] E. Ie, V. Jain, J. Wang, S. Narvekar, R. Agarwal, R. Wu, H.-T. Cheng, T. Chandra, and C. Boutilier. "SlateQ: A tractable decomposition for reinforcement learning with recommendation sets". In: *IJCAI*. International Joint Conferences on Artificial Intelligence Organization, 2019.

[Iof17] S. Ioffe. "Batch Renormalization: Towards Reducing Minibatch Dependence in Batch-Normalized Models". In: (2017). arXiv: 1702.03275 [cs.LG].

[IR10] A. Ilin and T. Raiko. "Practical Approaches to Principal Component Analysis in the Presence of Missing Values". In: *JMLR* 11 (2010), pp. 1957–2000.

[IS15] S. Ioffe and C. Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *ICML*. 2015, pp. 448–456.

[Isc+19] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. "Label Propagation for Deep Semi-supervised Learning". In: *CVPR*. 2019.

[Izm+18] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson. "Averaging Weights Leads to Wider Optima and Better Generalization". In: *UAI*. 2018.

[Izm+20] P. Izmailov, P. Kirichenko, M. Finzi, and A. G. Wilson. "Semi-supervised learning with normalizing flows". In: *ICML*. 2020, pp. 4615–4630.

[Jac+91] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton. "Adaptive mixtures of local experts". In: *Neural Computation* (1991).

[JAFF16] J. Johnson, A. Alahi, and L. Fei-Fei. "Perceptual Losses for Real-Time Style Transfer and Super-Resolution". In: *ECCV*. 2016.

[Jan18] E. Jang. *Normalizing Flows Tutorial*. https://blog.evjang.com/2018/01/nf1.html. 2018.

[Jay03] E. T. Jaynes. *Probability theory: the logic of science*. Cambridge university press, 2003.

[Jay76] E. T. Jaynes. "Confidence intervals vs Bayesian intervals". In: *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, vol II*. Ed. by W. L. Harper and C. A. Hooker. Reidel Publishing Co., 1976.

[JD88] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[JDJ17] J. Johnson, M. Douze, and H. Jégou. "Billion-scale similarity search with GPUs". In: (2017). arXiv: 1702.08734 [cs.CV].

[Jef61] H. Jeffreys. *Theory of Probability*. Oxford, 1961.

[Jef73] H. Jeffreys. *Scientific Inference*. Third edition. Cambridge, 1973.

[JH04] H. Jaeger and H. Haas. "Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication". In: *Science* 304.5667 (2004).

[JHG00] N. Japkowicz, S. Hanson, and M. Gluck. "Nonlinear autoassociation is not equivalent to PCA". In: *Neural Computation* 12 (2000), pp. 531–545.

[Jia+20] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio. "Fantastic Generalization Measures and Where to Find Them". In: *ICLR*. 2020.

[Jin+17] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song. "Neural Style Transfer: A Review". In: *arXiv [cs.CV]* (2017).

[JJ94] M. I. Jordan and R. A. Jacobs. "Hierarchical mixtures of experts and the EM algorithm". In: *Neural Computation* 6 (1994), pp. 181–214.

[JK13] A. Jern and C. Kemp. "A probabilistic account of exemplar and category generation". en. In: *Cogn. Psychol.* 66.1 (2013), pp. 85–125.

[JM08] D. Jurafsky and J. H. Martin. *Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd edition. Prentice-Hall, 2008.

[JM20] D. Jurafsky and J. H. Martin. *Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recog-*

nition (Third Edition). Draft of 3rd edition. 2020.

[JN04] S. Jain and R. M. Neal. "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model". In: *Journal of Computational and Graphical Statistics* 13.1 (2004), pp. 158–182.

[Jor19] M. Jordan. "Artificial Intelligence — The Revolution Hasn't Happened Yet". In: *Harvard Data Science Review* 1.1 (2019).

[JT19] L. Jing and Y. Tian. "Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey". In: (2019). arXiv: 1902.06162 [cs.CV].

[Jun+19] W. Jung, D. Jung, B. Kim, S. Lee, W. Rhee, and J. Anh. "Restructuring Batch Normalization to Accelerate CNN Training". In: *SysML*. 2019.

[JW19] S. Jain and B. C. Wallace. "Attention is not Explanation". In: *NAACL*. 2019.

[JZ13] R. Johnson and T. Zhang. "Accelerating Stochastic Gradient Descent using Predictive Variance Reduction". In: *NIPS*. Curran Associates, Inc., 2013, pp. 315–323.

[JZS15] R. Jozefowicz, W. Zaremba, and I. Sutskever. "An Empirical Exploration of Recurrent Network Architectures". In: *ICML*. 2015, pp. 2342–2350.

[KAG19] A. Kirsch, J. van Amersfoort, and Y. Gal. "BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning". In: *NIPS*. 2019.

[Kai58] H. Kaiser. "The varimax criterion for analytic rotation in factor analysis". In: *Psychometrika* 23.3 (1958).

[Kan+12] E. Kandel, J. Schwartz, T. Jessell, S. Siegelbaum, and A. Hudspeth, eds. *Principles of Neural Science*. Fifth Edition. 2012.

[Kan+20] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis. "Decoupling Representation and Classifier for Long-Tailed Recognition". In: *ICLR*. 2020.

[Kap16] J. Kaplan. *Artificial Intelligence: What Everyone Needs to Know*. en. 1st ed. Oxford University Press, 2016.

[Kap+20] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. "Scaling laws for neural language models". In: *arXiv [cs.LG]* (Jan. 2020).

[Kat+20] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. "Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention". In: *ICML*. 2020.

[KB15] D. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *ICLR*. 2015.

[KB19] M. Kaya and H. S. Bilge. "Deep Metric Learning: A Survey". en. In: *Symmetry* 11.9 (2019), p. 1066.

[KBV09] Y. Koren, R. Bell, and C. Volinsky. "Matrix factorization techniques for recommender systems". In: *IEEE Computer* 42.8 (2009), pp. 30–37.

[KD09] A. D. Kiureghian and O. Ditlevsen. "Aleatory or epistemic? Does it matter?" In: *Structural Safety* 31.2 (2009), pp. 105–112.

[Kem+06] C. Kemp, J. Tenenbaum, T. Y. T. Griffiths and, and N. Ueda. "Learning systems of concepts with an infinite relational model". In: *AAAI*. 2006.

[KF05] H. Kuck and N. de Freitas. "Learning about individuals from group statistics". In: *UAI*. 2005.

[KG05] A. Krause and C. Guestrin. "Near-optimal value of information in graphical models". In: *UAI*. 2005.

[KG17] A. Kendall and Y. Gal. "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" In: *NIPS*. Curran Associates, Inc., 2017, pp. 5574–5584.

[KGS20] J. von Kügelgen, L. Gresele, and B. Schölkopf. "Simpson's paradox in Covid-19 case fatality rates: a mediation analysis of age-related causal effects". In: (2020). arXiv: 2005.07180 [stat.AP].

[KH09] A Krizhevsky and G Hinton. *Learning multiple layers of features from tiny images*. Tech. rep. U. Toronto, 2009.

[KH19] D. Krotov and J. J. Hopfield. "Unsupervised learning by competing hidden units". en. In: *PNAS* 116.16 (2019), pp. 7723–7731.

[Kha+10] M. E. Khan, B. Marlin, G. Bouchard, and K. P. Murphy. "Variational bounds for mixed-data factor analysis". In: *NIPS*. 2010.

[Kha+20] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi. "A Survey of the Recent Architectures of Deep Convolutional Neural Networks". In: *AI Review* (2020).

[KHB07] A. Kapoor, E. Horvitz, and S. Basu. "Selective Supervision: Guiding Supervised Learning with Decision-Theoretic Active Learning". In: *IJCAI*. 2007.

[KHW19] W. Kool, H. van Hoof, and M. Welling. "Stochastic Beams and Where to Find Them: The Gumbel-Top-k Trick for Sampling Sequences Without Replacement". In: *ICML*. 2019.

[Kim14] Y. Kim. “Convolutional Neural Networks for Sentence Classification”. In: *EMNLP*. 2014.

[Kim19] D. H. Kim. *Survey of Deep Metric Learning*. 2019.

[Kin+14] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling. “Semi-Supervised Learning with Deep Generative Models”. In: *NIPS*. 2014.

[Kir+19] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. “Panoptic Segmentation”. In: *CVPR*. 2019.

[KJ16] L Kang and V Joseph. “Kernel Approximation: From Regression to Interpolation”. In: *SIAM/ASA J. Uncertainty Quantification* 4.1 (2016), pp. 112–129.

[KJ95] J. Karhunen and J. Joutsensalo. “Generalizations of principal component analysis, optimization problems, and neural networks”. In: *Neural Networks* 8.4 (1995), pp. 549–562.

[KJM19] N. M. Kriege, F. D. Johansson, and C. Morris. “A Survey on Graph Kernels”. In: (2019). arXiv: 1903.11835 [cs.LG].

[KKH20] I. Khemakhem, D. P. Kingma, and A. Hyvärinen. “Variational Autoencoders and Nonlinear ICA: A Unifying Framework”. In: *AISTATS*. 2020.

[KKL20] N. Kitaev, L. Kaiser, and A. Levskaya. “Reformer: The Efficient Transformer”. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[KKS20] F. Kunstner, R. Kumar, and M. Schmidt. “Homeomorphic-Invariance of EM: Non-Asymptotic Convergence in KL Divergence for Exponential Families via Mirror Descent”. In: (2020). arXiv: 2011.01170 [cs.LG].

[KL17] J. K. Kruschke and T. M. Liddell. “The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective”. In: *Psychon. Bull. Rev.* (2017).

[KL21] W. M. Kouw and M. Loog. “A review of domain adaptation without target labels”. en. In: *IEEE PAMI* (2021).

[Kla+17] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. “Self-Normalizing Neural Networks”. In: *NIPS*. 2017.

[Kle02] J. Kleinberg. “An Impossibility Theorem for Clustering”. In: *NIPS*. 2002.

[Kle+11] A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan. *A scalable bootstrap for massive data*. Tech. rep. UC Berkeley, 2011.

[Kle13] P. N. Klein. *Coding the Matrix: Linear Algebra through Applications to Computer Science*. en. 1 edition. Newtonian Press, 2013.

[KLQ95] C. Ko, J. Lee, and M. Queyranne. “An exact algorithm for maximum entropy sampling”. In: *Operations Research* 43 (1995), 684–691.

[Kok17] I. Kokkinos. “UberNet: Training a Universal Convolutional Neural Network for Low-, Mid-, and High-Level Vision Using Diverse Datasets and Limited Memory”. In: *CVPR*. Vol. 2. 2017, p. 8.

[Kol+19] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby. “Large Scale Learning of General Visual Representations for Transfer”. In: (2019). arXiv: 1912.11370 [cs.CV].

[Kol+20] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby. “Large Scale Learning of General Visual Representations for Transfer”. In: *ECCV*. 2020.

[Kon20] M. Konnikova. *The Biggest Bluff: How I Learned to Pay Attention, Master Myself, and Win*. en. Penguin Press, 2020.

[Kor09] Y. Koren. *The BellKor Solution to the Netflix Grand Prize*. Tech. rep. Yahoo! Research, 2009.

[KR19] M. Kearns and A. Roth. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. en. Oxford University Press, 2019.

[KR87] L. Kaufman and P. Rousseeuw. “Clustering by means of Medoids”. In: *Statistical Data Analysis Based on the L1-norm and Related Methods*. Ed. by Y. Dodge. North-Holland, 1987, 405–416.

[KR90] L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990.

[Kri+05] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink. “Learning sparse Bayesian classifiers: multi-class formulation, fast algorithms, and generalization bounds”. In: *IEEE Transaction on Pattern Analysis and Machine Intelligence* (2005).

[Kru13] J. K. Kruschke. “Bayesian estimation supersedes the t test”. In: *J. Experimental Psychology: General* 142.2 (2013), pp. 573–603.

[Kru15] J. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS and STAN*. Second edition. Academic Press, 2015.

[KS15] H. Kaya and A. A. Salah. “Adaptive Mixtures of Factor Analyzers”. In: (2015). arXiv: 1507.02801 [stat.ML].

[KSG04] A. Kraskov, H. Stögbauer, and P. Grassberger. “Estimating mutual information”. en. In: *Phys. Rev. E Stat. Non-*

lin. Soft Matter Phys. 69.6 Pt 2 (2004), p. 066138.

[KSH12] A. Krizhevsky, I. Sutskever, and G. Hinton. "Imagenet classification with deep convolutional neural networks". In: *NIPS*. 2012.

[KSJ09] I. Konstas, V. Stathopoulos, and J. M. Jose. "On social networks and collaborative recommendation". In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 2009, pp. 195–202.

[KST82] D. Kahneman, P. Slovic, and A. Tversky, eds. *Judgment under uncertainty: Heuristics and biases*. Cambridge, 1982.

[KTB11] D. P. Kroese, T. Taimre, and Z. I. Botev. *Handbook of Monte Carlo Methods*. en. 1 edition. Wiley, 2011.

[Kua+09] P. Kuan, G. Pan, J. A. Thomson, R. Stewart, and S. Keles. *A hierarchical semi-Markov model for detecting enrichment with application to ChIP-Seq experiments*. Tech. rep. U. Wisconsin, 2009.

[Kul13] B. Kulis. "Metric Learning: A Survey". In: *Foundations and Trends in Machine Learning* 5.4 (2013), pp. 287–364.

[KV94] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994.

[KVK10] A. Klami, S. Virtanen, and S. Kaski. "Bayesian exponential family projections for coupled data sources". In: *UAI*. 2010.

[KW14] D. P. Kingma and M. Welling. "Auto-encoding variational Bayes". In: *ICLR*. 2014.

[KW16a] T. N. Kipf and M. Welling. "Semi-supervised classification with graph convolutional networks". In: *arXiv preprint arXiv:1609.02907* (2016).

[KW16b] T. N. Kipf and M. Welling. "Variational graph auto-encoders". In: *arXiv preprint arXiv:1611.07308* (2016).

[KW19a] D. P. Kingma and M. Welling. "An Introduction to Variational Autoencoders". In: *Foundations and Trends in Machine Learning* 12.4 (2019), pp. 307–392.

[KW19b] M. J. Kochenderfer and T. A. Wheeler. *Algorithms for Optimization*. en. The MIT Press, 2019.

[KWW22] M. J. Kochenderfer, T. A. Wheeler, and K. Wray. *Algorithms for Decision Making*. The MIT Press, 2022.

[Kyu+10] M. Kyung, J. Gill, M. Ghosh, and G. Casella. "Penalized Regression, Standard Errors and Bayesian Lassos". In: *Bayesian Analysis* 5.2 (2010), pp. 369–412.

[LA16] S. Laine and T. Aila. "Temporal ensembling for semi-supervised learning". In: *arXiv preprint arXiv:1610.02242* (2016).

[Lak+17] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. "Building Machines That Learn and Think Like People". en. In: *Behav. Brain Sci.* (2017), pp. 1–101.

[Lam18] B. Lambert. *A Student's Guide to Bayesian Statistics*. en. 1st ed. SAGE Publications Ltd, 2018.

[Lam25] N. Lambert. *Reinforcement Learning from Human Feedback*. 2025.

[Law12] N. D. Lawrence. "A Unifying Probabilistic Perspective for Spectral Dimensionality Reduction: Insights and New Models". In: *JMLR* 13.May (2012), pp. 1609–1638.

[LBM06] J. Lasserre, C. Bishop, and T. Minka. "Principled Hybrids of Generative and Discriminative Models". In: *CVPR*. 2006.

[LBS19] Y. Li, J. Bradshaw, and Y. Sharma. "Are Generative Classifiers More Robust to Adversarial Attacks?" In: *ICML*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 3804–3814.

[LeC18] Y. LeCun. *Self-supervised learning: could machines learn like humans?* 2018.

[LeC+98] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-Based Learning Applied to Document Recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[Lee13] D.-H. Lee. "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks". In: *ICML Workshop on Challenges in Representation Learning*. 2013.

[Lee+13] J. Lee, S. Kim, G. Lebanon, and Y. Singer. "Local Low-Rank Matrix Approximation". In: *ICML*. Vol. 28. Proceedings of Machine Learning Research. PMLR, 2013, pp. 82–90.

[Lee+19] J. Lee, Y. Lee, J. Kim, A. R. Kosiorek, S. Choi, and Y. W. Teh. "Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks". In: *ICML*. 2019.

[Lee77] J. de Leeuw. "Applications of Convex Analysis to Multidimensional Scaling". In: *Recent Developments in Statistics*. Ed. by J. R. Barra, F Brodeau, G Romier, and B Van Cutsem. 1977.

[Lep+21] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen. "GShard: Scaling Giant Models with Conditional Computation

and Automatic Sharding". In: *ICLR*. 2021.

[LG14] O. Levy and Y. Goldberg. "Neural Word Embedding as Implicit Matrix Factorization". In: *NIPS*. 2014.

[LH17] I. Loshchilov and F. Hutter. "SGDR: Stochastic Gradient Descent with Warm Restarts". In: *ICLR*. 2017.

[Li+15] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel. "Gated graph sequence neural networks". In: *arXiv preprint arXiv:1511.05493* (2015).

[Li+17] A. Li, A. Jabri, A. Joulin, and L. van der Maaten. "Learning Visual N-Grams from Web Data". In: *ICCV*. 2017.

[Lia20] S. M. Liao, ed. *Ethics of Artificial Intelligence*. en. 1st ed. Oxford University Press, 2020.

[Lim+19] S. Lim, I. Kim, T. Kim, C. Kim, and S. Kim. "Fast AutoAugment". In: (2019). arXiv: 1905.00397 [cs.LG].

[Lin06] D. Lindley. *Understanding Uncertainty*. Wiley, 2006.

[Lin+21] T. Lin, Y. Wang, X. Liu, and X. Qiu. "A Survey of Transformers". In: (2021). arXiv: 2106.04554 [cs.LG].

[Lin56] D. Lindley. "On a measure of the information provided by an experiment". In: *The Annals of Math. Stat.* (1956), 986–1005.

[Liu01] J. Liu. *Monte Carlo Strategies in Scientific Computation*. Springer, 2001.

[Liu+16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. "SSD: Single Shot MultiBox Detector". In: *ECCV*. 2016.

[Liu+18a] H. Liu, Y.-S. Ong, X. Shen, and J. Cai. "When Gaussian Process Meets Big Data: A Review of Scalable GPs". In: (2018). arXiv: 1807.01065 [stat.ML].

[Liu+18b] L. Liu, X. Liu, C.-J. Hsieh, and D. Tao. "Stochastic Second-order Methods for Non-convex Optimization with Inexact Hessian and Gradient". In: (2018). arXiv: 1809.09853 [math.OC].

[Liu+20] F. Liu, X. Huang, Y. Chen, and J. A. K. Suykens. "Random Features for Kernel Approximation: A Survey on Algorithms, Theory, and Beyond". In: (2020). arXiv: 2004.11154 [stat.ML].

[Liu+22] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. "A ConvNet for the 2020s". In: (2022). arXiv: 2201.03545 [cs.CV].

[LKB20] Q. Liu, M. J. Kusner, and P. Blunsom. "A Survey on Contextual Embeddings". In: (2020). arXiv: 2003.07278 [cs.CL].

[Llo82] S Lloyd. "Least squares quantization in PCM". In: *IEEE Trans. Inf. Theory* 28.2 (1982), pp. 129–137.

[LLT89] K. Lange, R. Little, and J. Taylor. "Robust Statistical Modeling Using the T Distribution". In: *JASA* 84.408 (1989), pp. 881–896.

[LM04] E. Learned-Miller. *Hyperspacings and the estimation of information theoretic quantities*. Tech. rep. 04-104. U. Mass. Amherst Comp. Sci. Dept, 2004.

[LM86] R. Larsen and M. Marx. *An introduction to mathematical statistics and its applications*. Prentice Hall, 1986.

[LN81] D. V. Lindley and M. R. Novick. "The Role of Exchangeability in Inference". en. In: *Annals of Statistics* 9.1 (1981), pp. 45–58.

[Loa00] C. F. V. Loan. "The ubiquitous Kronecker product". In: *J. Comput. Appl. Math.* 123.1 (2000), pp. 85–100.

[Lod+02] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. "Text classification using string kernels". en. In: *J. Mach. Learn. Res.* (2002).

[LPM15] M.-T. Luong, H. Pham, and C. D. Manning. "Effective Approaches to Attention-based Neural Machine Translation". In: *EMNLP*. 2015.

[LR87] R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley and Son, 1987.

[LRU14] J. Leskovec, A. Rajaraman, and J. Ullman. *Mining of massive datasets*. Cambridge, 2014.

[LS10] P. Long and R. Servedio. "Random classification noise beats all convex potential boosters". In: *JMLR* 78.3 (2010), pp. 287–304.

[LS19a] S. Lattanzi and C. Sohler. "A Better k-means++ Algorithm via Local Search". In: *ICML*. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 3662–3671.

[LS19b] Z. C. Lipton and J. Steinhardt. "Troubling Trends in Machine Learning Scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research". In: *The Queue* 17.1 (2019), pp. 45–77.

[LSS13] Q. Le, T. Sarlos, and A. Smola. "Fastfood - Computing Hilbert Space Expansions in loglinear time". In: *ICML*. Vol. 28. Proceedings of Machine Learning Research. PMLR, 2013, pp. 244–252.

[LSY19] H. Liu, K. Simonyan, and Y. Yang. "DARTS: Differentiable Architecture Search". In: *ICLR*. 2019.

[Lu+19] L. Lu, Y. Shin, Y. Su, and G. E. Karniadakis. "Dying ReLU and Initialization: Theory and Numerical Examples". In: (2019). arXiv: 1903.06733 [stat.ML].

[Luo16] M.-T. Luong. “Neural machine translation”. PhD thesis. Stanford Dept. Comp. Sci., 2016.

[Luo+19] P. Luo, X. Wang, W. Shao, and Z. Peng. “Towards Understanding Regularization in Batch Normalization”. In: *ICLR*. 2019.

[LUW17] C. Louizos, K. Ullrich, and M. Welling. “Bayesian Compression for Deep Learning”. In: *NIPS*. 2017.

[Lux07] U. von Luxburg. “A tutorial on spectral clustering”. In: *Statistics and Computing* 17.4 (2007), pp. 395–416.

[LW04a] O. Ledoit and M. Wolf. “A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices”. In: *J. of Multivariate Analysis* 88.2 (2004), pp. 365–411.

[LW04b] O. Ledoit and M. Wolf. “Honey, I Shrunk the Sample Covariance Matrix”. In: *J. of Portfolio Management* 31.1 (2004).

[LW04c] H. Lopes and M. West. “Bayesian model assessment in factor analysis”. In: *Statisica Sinica* 14 (2004), pp. 41–67.

[LW16] C. Li and M. Wand. “Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks”. In: *ECCV*. 2016.

[LWG12] U. von Luxburg, R. Williamson, and I. Guyon. “Clustering: science or art?” In: *Workshop on Unsupervised and Transfer Learning*. 2012.

[LXW19] X. Liu, Q. Xu, and N. Wang. “A survey on deep neural network-based image captioning”. In: *The Visual Computer* 35.3 (2019), pp. 445–470.

[Lyu+20] X.-K. Lyu, Y. Xu, X.-F. Zhao, X.-N. Zuo, and C.-P. Hu. “Beyond psychology: prevalence of p value and confidence interval misinterpretation across different fields”. In: *Journal of Pacific Rim Psychology* 14 (2020).

[MA10] I. Murray and R. P. Adams. “Slice sampling covariance hyperparameters of latent Gaussian models”. In: *NIPS*. 2010, pp. 1732–1740.

[MA+17] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh. “No Fuss Distance Metric Learning using Proxies”. In: *ICCV*. 2017.

[Maa+11] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. “Learning Word Vectors for Sentiment Analysis”. In: *Proc. ACL*. 2011, pp. 142–150.

[Maa14] L. van der Maaten. “Accelerating t-SNE using Tree-Based Algorithms”. In: *JMLR* (2014).

[Mac03] D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[Mac09] L. W. Mackey. “Deflation Methods for Sparse PCA”. In: *NIPS*. 2009.

[Mac67] J MacQueen. “Some methods for classification and analysis of multivariate observations”. en. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. The Regents of the University of California, 1967.

[Mac95] D. MacKay. “Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks”. In: *Network: Computation in Neural Systems* 6.3 (1995), pp. 469–505.

[Mad+20] A. Madani, B. McCann, N. Naik, N. S. Keskar, N. Anand, R. R. Eguchi, P.-S. Huang, and R. Socher. “ProGen: Language Modeling for Protein Generation”. en. 2020.

[Mah07] R. P. S. Mahler. *Statistical Multisource-Multitarget Information Fusion*. Artech House, Inc., 2007.

[Mah13] R Mahler. “Statistics 102 for Multisource-Multitarget Detection and Tracking”. In: *IEEE J. Sel. Top. Signal Process.* 7.3 (2013), pp. 376–389.

[Mah+18] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. “Exploring the Limits of Weakly Supervised Pretraining”. In: (2018). arXiv: 1805.00932 [cs.CV].

[Mai15] J Mairal. “Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning”. In: *SIAM J. Optim.* 25.2 (2015), pp. 829–855.

[Mak+19] D. Makowski, M. S. Ben-Shachar, S. H. A. Chen, and D. Lüdecke. “Indices of Effect Existence and Significance in the Bayesian Framework”. en. In: *Front. Psychol.* 10 (2019), p. 2767.

[Mal99] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[Man+16] V. Mansinghka, P. Shafto, E. Jonas, C. Petschulat, M. Gasner, and J. Tenenbaum. “Crosscat: A Fully Bayesian, Nonparametric Method For Analyzing Heterogeneous, High-dimensional Data.” In: *JMLR* 17 (2016).

[Mar06] H. Markram. “The blue brain project”. en. In: *Nat. Rev. Neurosci.* 7.2 (2006), pp. 153–160.

[Mar08] B. Marlin. “Missing Data Problems in Machine Learning”. PhD thesis. U. Toronto, 2008.

[Mar+11] B. M. Marlin, R. S. Zemel, S. T. Roweis, and M. Slaney. “Recommender Systems,

Missing Data and Statistical Model Estimation". In: *IJCAI*. 2011.

[Mar18] O. Martin. *Bayesian analysis with Python*. Packt, 2018.

[Mar72] G. Marsaglia. "Choosing a Point from the Surface of a Sphere". en. In: *Ann. Math. Stat.* 43.2 (1972), pp. 645–646.

[Mas+00] L. Mason, J. Baxter, P. L. Bartlett, and M. R. Frean. "Boosting Algorithms as Gradient Descent". In: *NIPS*. 2000, pp. 512–518.

[Mas+15] J. Masci, D. Boscaini, M. Bronstein, and P. Vandergheynst. "Geodesic convolutional neural networks on riemannian manifolds". In: *Proceedings of the IEEE international conference on computer vision workshops*. 2015, pp. 37–45.

[Mat00] R. Matthews. "Storks Deliver Babies (p = 0.008)". In: *Teach. Stat.* 22.2 (2000), pp. 36–38.

[Mat98] R. Matthews. *Bayesian Critique of Statistics in Health: The Great Health Hoax*. 1998.

[MAV17] D. Molchanov, A. Ashukha, and D. Vetrov. "Variational Dropout Sparsifies Deep Neural Networks". In: *ICML*. 2017.

[MB05] F. Morin and Y. Bengio. "Hierarchical Probabilistic Neural Network Language Model". In: *AISTATS*. 2005.

[MB06] N. Meinshausen and P. Buhlmann. "High dimensional graphs and variable selection with the lasso". In: *The Annals of Statistics* 34 (2006), pp. 1436–1462.

[MBL20] K. Musgrave, S. Belongie, and S.-N. Lim. "A Metric Learning Reality Check". In: *ECCV*. 2020.

[McE20] R. McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan (2nd edition)*. en. Chapman and Hall/CRC, 2020.

[McL75] G. J. McLachlan. "Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis". In: *Journal of the American Statistical Association* 70.350 (1975), pp. 365–369.

[MD97] X. L. Meng and D. van Dyk. "The EM algorithm — an old folk song sung to a fast new tune (with Discussion)". In: *J. Royal Stat. Soc. B* 59 (1997), pp. 511–567.

[ME14] S. Masoudnia and R. Ebrahimpour. "Mixture of experts: a literature survey". In: *Artificial Intelligence Review* 42.2 (2014), pp. 275–293.

[Mei01] M. Meila. "A random walks view of spectral segmentation". In: *AISTATS*. 2001.

[Mei05] M. Meila. "Comparing clusterings: an axiomatic view". In: *ICML*. 2005.

[Men+12] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. "Metric Learning for Large Scale Image Classification: Generalizing to New Classes at Near-Zero Cost". In: *ECCV*. Springer Berlin Heidelberg, 2012, pp. 488–501.

[Met21] C. Metz. *Genius Makers: The Mavericks Who Brought AI to Google, Facebook, and the World*. en. Dutton, 2021.

[MF17] J. Matejka and G. Fitzmaurice. "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2017, pp. 1290–1294.

[MFR20] G. M. Martin, D. T. Frazier, and C. P. Robert. "Computing Bayes: Bayesian Computation from 1763 to the 21st Century". In: (2020). arXiv: 2004 . 06425 [stat.CO].

[MG05] I. Murray and Z. Ghahramani. *A note on the evidence and Bayesian Occam's razor*. Tech. rep. Gatsby, 2005.

[MH07] A. Mnih and G. Hinton. "Three new graphical models for statistical language modelling". en. In: *ICML*. 2007.

[MH08] L. v. d. Maaten and G. Hinton. "Visualizing Data using t-SNE". In: *JMLR* 9.Nov (2008), pp. 2579–2605.

[MHM18] L. McInnes, J. Healy, and J. Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: (2018). arXiv: 1802 . 03426 [stat.ML].

[MHN13] A. L. Maas, A. Y. Hannun, and A. Y. Ng. "Rectifier Nonlinearities Improve Neural Network Acoustic Models". In: *ICML*. Vol. 28. 2013.

[Mik+13a] T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient Estimation of Word Representations in Vector Space". In: *ICLR*. 2013.

[Mik+13b] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. "Distributed Representations of Words and Phrases and their Compositionality". In: *NIPS*. 2013.

[Mik+13c] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. "Distributed representations of words and phrases and their compositionality". In: *NIPS*. 2013, pp. 3111–3119.

[Min00] T. Minka. *Bayesian model averaging is not model combination*. Tech. rep. MIT Media Lab, 2000.

[Min+09] M. Mintz, S. Bills, R. Snow, and D. Jurafksy. "Distant supervision for relation extraction without labeled data". In:

Prof. Conf. Recent Advances in NLP. 2009.

[Mit97] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[Miy+18] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii. "Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning". In: *IEEE PAMI* (2018).

[MK97] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 1997.

[MKH19] R. Müller, S. Kornblith, and G. E. Hinton. "When does label smoothing help?" In: *NIPS*. 2019, pp. 4694–4703.

[MKL11] O. Martin, R. Kumar, and J. Lao. *Bayesian Modeling and Computation in Python*. CRC Press, 2011.

[MKS21] K. Murphy, A. Kumar, and S. Serghiou. "Risk score learning for COVID-19 contact tracing apps". In: *Machine Learning for Healthcare*. 2021.

[MM16] D. Mishkin and J. Matas. "All you need is a good init". In: *ICLR*. 2016.

[MN89] P. McCullagh and J. Nelder. *Generalized linear models*. 2nd edition. Chapman and Hall, 1989.

[MNM02] W. Maass, T. Natschlaeger, and H. Markram. "Real-time computing without stable states: A new framework for neural computation based on perturbations". In: *Neural Computation* 14.11 (2002), 2531—2560.

[MO04] S. C. Madeira and A. L. Oliveira. "Biclustering Algorithms for Biological Data Analysis: A Survey". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1.1 (2004), pp. 24–45.

[Mol04] C. Moler. *Numerical Computing with MATLAB*. SIAM, 2004.

[Mon+14] G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio. "On the Number of Linear Regions of Deep Neural Networks". In: *NIPS*. 2014.

[Mon+17] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein. "Geometric deep learning on graphs and manifolds using mixture model cnns". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 5115–5124.

[Mon+19] N. Monath, A. Kobren, A. Krishnamurthy, M. R. Glass, and A. McCallum. "Scalable Hierarchical Clustering with Tree Grafting". In: *KDD*. KDD '19. Association for Computing Machinery, 2019, pp. 1438–1448.

[Mon+21] N. Monath et al. "Scalable Bottom-Up Hierarchical Clustering". In: *KDD*. 2021.

[Mor+16] R. D. Morey, R. Hoekstra, J. N. Rouder, M. D. Lee, and E.-J. Wagenmakers. "The fallacy of placing confidence in confidence intervals". en. In: *Psychon. Bull. Rev.* 23.1 (2016), pp. 103–123.

[MOT15] A. Mordvintsev, C. Olah, and M. Tyka. *Inceptionism: Going Deeper into Neural Networks*. https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html. Accessed: NA-NA-NA. 2015.

[MP43] W. McCulloch and W. Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *Bulletin of Mathematical Biophysics* 5 (1943), pp. 115–137.

[MP69] M. Minsky and S. Papert. *Perceptrons*. MIT Press, 1969.

[MRS08] C. Manning, P. Raghavan, and H. Schuetze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[MS11] D. Mayo and A. Spanos. "Error Statistics". In: *Handbook of Philosophy of Science*. Ed. by P. S. Bandyopadhyay and M. R. Forster. 2011.

[Muk+19] B. Mukhoty, G. Gopakumar, P. Jain, and P. Kar. "Globally-convergent Iteratively Reweighted Least Squares for Robust Regression Problems". In: *AISTATS*. 2019, pp. 313–322.

[Mur23] K. P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[MV15] A Mahendran and A Vedaldi. "Understanding deep image representations by inverting them". In: *CVPR*. 2015, pp. 5188–5196.

[MV16] A. Mahendran and A. Vedaldi. "Visualizing Deep Convolutional Neural Networks Using Natural Pre-images". In: *Intl. J. Computer Vision* (2016), pp. 1–23.

[MWK16] A. H. Marblestone, G. Wayne, and K. P. Kording. "Toward an Integration of Deep Learning and Neuroscience". en. In: *Front. Comput. Neurosci.* 10 (2016), p. 94.

[MWP98] B Moghaddam, W Wahid, and A Pentland. "Beyond eigenfaces: probabilistic matching for face recognition". In: *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. 1998, pp. 30–35.

[Nad+19] S. Naderi, K. He, R. Aghajani, S. Sclaroff, and P. Felzenszwalb. "Generalized Majorization-Minimization". In: *ICML*. 2019.

[Nak+19] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever. "Deep

Double Descent: Where Bigger Models and More Data Hurt". In: (2019). arXiv: 1912.02292 [cs.LG].

[NAM21] C. G. Northcutt, A. Athalye, and J. Mueller. "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks". In: *NeurIPS Track on Datasets and Benchmarks*. Mar. 2021.

[Nea96] R. Neal. *Bayesian learning for neural networks*. Springer, 1996.

[Nes04] Y. Nesterov. *Introductory Lectures on Convex Optimization. A basic course*. Kluwer, 2004.

[Neu04] A. Neumaier. "Complete search in continuous global optimization and constraint satisfaction". In: *Acta Numer.* 13 (2004), pp. 271–369.

[Neu17] G. Neubig. "Neural Machine Translation and Sequence-to-sequence Models: A Tutorial". In: (2017). arXiv: 1703 . 01619 [cs.CL].

[Ngu+17] A. Nguyen, J. Yosinski, Y. Bengio, A. Dosovitskiy, and J. Clune. "Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space". In: *CVPR*. 2017.

[NHLS19] E. Nalisnick, J. M. Hernández-Lobato, and P. Smyth. "Dropout as a Structured Shrinkage Prior". In: *ICML*. 2019.

[Nic+15] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. "A Review of Relational Machine Learning for Knowledge Graphs". In: *Proc. IEEE* (2015).

[Niu+11] F. Niu, B. Recht, C. Re, and S. J. Wright. "HOGWILD!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent". In: *NIPS*. 2011.

[NJ02] A. Y. Ng and M. I. Jordan. "On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes". In: *NIPS-14*. 2002.

[NJW01] A. Ng, M. Jordan, and Y. Weiss. "On Spectral Clustering: Analysis and an algorithm". In: *NIPS*. 2001.

[NK17] M. Nickel and D. Kiela. "Poincaré embeddings for learning hierarchical representations". In: *Advances in neural information processing systems*. 2017, pp. 6338–6347.

[NK18] M. Nickel and D. Kiela. "Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry". In: *International Conference on Machine Learning*. 2018, pp. 3779–3788.

[NK24] A. Narayanan and S. Kapoor. *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*. en. 2024.

[NMC05] A. Niculescu-Mizil and R. Caruana. "Predicting Good Probabilities with Supervised Learning". In: *ICML*. 2005.

[Nou+02] M. N. Nounou, B. R. Bakshi, P. K. Goel, and X. Shen. "Process modeling by Bayesian latent variable regression". In: *Am. Inst. Chemical Engineers Journal* 48.8 (2002), pp. 1775–1793.

[Nov62] A. Novikoff. "On convergence proofs on perceptrons". In: *Symp. on the Mathematical Theory of Automata* 12 (1962), pp. 615–622.

[NR18] G. Neu and L. Rosasco. "Iterate Averaging as Regularization for Stochastic Gradient Descent". In: *COLT*. 2018.

[NTL20] J. Nixon, D. Tran, and B. Lakshminarayanan. "Why aren't bootstrapped neural networks better?" In: *NIPS Workshop on "I can't believe it's not better"*. 2020.

[NW06] J. Nocedal and S. Wright. *Numerical Optimization*. Springer, 2006.

[Ode16] A. Odena. "Semi-supervised learning with generative adversarial networks". In: *arXiv preprint arXiv:1606.01583* (2016).

[OLV18] A. van den Oord, Y. Li, and O. Vinyals. "Representation Learning with Contrastive Predictive Coding". In: (2018). arXiv: 1807.03748 [cs.LG].

[OMS17] C. Olah, A. Mordvintsev, and L. Schubert. "Feature Visualization". In: *Distill* (2017).

[Oor+16] A. Van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. "WaveNet: A Generative Model for Raw Audio". In: (2016). arXiv: 1609.03499 [cs.SD].

[Oor+18] A. van den Oord et al. "Parallel WaveNet: Fast High-Fidelity Speech Synthesis". In: *ICML*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 3918–3926.

[Ope] OpenAI. *ChatGPT: Optimizing Language Models for Dialogue*. Blog.

[Ope23] OpenAI. *GPT4*. Tech. rep. 2023.

[OPK12] G. Ohloff, W. Pickenhagen, and P. Kraft. *Scent and Chemistry*. en. Wiley, 2012.

[OPT00a] M. R. Osborne, B. Presnell, and B. A. Turlach. "A new approach to variable selection in least squares problems". In: *IMA Journal of Numerical Analysis* 20.3 (2000), pp. 389–403.

[OPT00b] M. R. Osborne, B. Presnell, and B. A. Turlach. "On the lasso and its dual". In: *J. Computational and graphical statistics* 9 (2000), pp. 319–337.

[Ort+19] P. A. Ortega et al. “Meta-learning of Sequential Strategies”. In: (2019). arXiv: 1905.03030 [cs.LG].

[Osb16] I. Osband. “Risk versus Uncertainty in Deep Learning: Bayes, Bootstrap and the Dangers of Dropout”. In: *NIPS workshop on Bayesian deep learning*. 2016.

[OTJ07] G. Obozinski, B. Taskar, and M. I. Jordan. *Joint covariate selection for grouped classification*. Tech. rep. UC Berkeley, 2007.

[Ouy+22] L. Ouyang et al. “Training language models to follow instructions with human feedback”. In: (Mar. 2022). arXiv: 2203.02155 [cs.CL].

[Pai05] A. Pais. *Subtle Is the Lord: The Science and the Life of Albert Einstein*. en. Oxford University Press, 2005.

[Pan+15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. “Librispeech: an asr corpus based on public domain audio books”. In: *ICASSP*. IEEE. 2015, pp. 5206–5210.

[Pap+18] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy. “PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model”. In: *ECCV*. 2018, pp. 269–286.

[Par+16a] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit. “A Decomposable Attention Model for Natural Language Inference”. In: *EMNLP*. Association for Computational Linguistics, 2016, pp. 2249–2255.

[Par+16b] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit. “A Decomposable Attention Model for Natural Language Inference”. In: *EMNLP*. Association for Computational Linguistics, 2016, pp. 2249–2255.

[Par+18] N. Parmar, A. Vaswani, J. Uszkoreit, Ł. Kaiser, N. Shazeer, A. Ku, and D. Tran. “Image Transformer”. In: *ICLR*. 2018.

[PARS14] B. Perozzi, R. Al-Rfou, and S. Skiena. “Deepwalk: Online learning of social representations”. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2014, pp. 701–710.

[Pas14] R. Pascanu. “On Recurrent and Deep Neural Networks”. PhD thesis. U. Montreal, 2014.

[Pat12] A. Paterek. *Predicting movie ratings and recommender systems*. 2012.

[Pat+16] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. “Context Encoders: Feature Learning by Inpainting”. In: *CVPR*. 2016.

[Pau+20] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna. “Data and its (dis)contents: A survey of dataset development and use in machine learning research”. In: *NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-analyses (ML-RSA)*. 2020.

[PB+14] N. Parikh, S. Boyd, et al. “Proximal algorithms”. In: *Foundations and Trends in Optimization* 1.3 (2014), pp. 127–239.

[Pea18] J. Pearl. *Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution*. Tech. rep. UCLA, 2018.

[Pen+20] Z. Peng, W. Huang, M. Luo, Q. Zheng, Y. Rong, T. Xu, and J. Huang. “Graph Representation Learning via Graphical Mutual Information Maximization”. In: *Proceedings of The Web Conference*. 2020.

[Per+17] B. Perozzi, V. Kulkarni, H. Chen, and S. Skiena. “Don’t Walk, Skip! Online Learning of Multi-Scale Network Embeddings”. In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ASONAM ’17. Association for Computing Machinery, 2017, 258–265.

[Pet13] J. Peters. *When Ice Cream Sales Rise, So Do Homicides. Coincidence, or Will Your Next Cone Murder You?* https://slate.com/news-and-politics/2013/07/warm-weather-homicide-rates-when-ice-cream-sales-rise-homicides-rise-coincidence.html. Accessed: 2020-5-20. 2013.

[Pet+18] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. “Deep contextualized word representations”. In: *NAACL*. 2018.

[Pey20] G. Peyre. “Course notes on Optimization for Machine Learning”. 2020.

[PH18] T. Parr and J. Howard. “The Matrix Calculus You Need For Deep Learning”. In: (2018). arXiv: 1802.01528 [cs.LG].

[Pin88] F. J. Pineda. “Generalization of back propagation to recurrent and higher order neural networks”. In: *Neural information processing systems*. 1988, pp. 602–611.

[Piz01] Z Pizlo. “Perception viewed as an inverse problem”. en. In: *Vision Res.* 41.24 (2001), pp. 3145–3161.

[PJ09] H.-S. Park and C.-H. Jun. “A simple and fast algorithm for K-medoids clustering”. In: *Expert Systems with Applciations* 36.2, Part 2 (2009), pp. 3336–3341.

[PJ92] B Polyak and A Juditsky. “Acceleration of Stochastic Approximation by Averaging”. In: *SIAM J. Control Optim.* 30.4 (1992), pp. 838–855.

[Pla00] J. Platt. “Probabilities for SV machines”. In: *Advances in Large Margin Classi-*

fiers. Ed. by A. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans. MIT Press, 2000.

[Pla98] J. Platt. "Using analytic QP and sparseness to speed training of support vector machines". In: *NIPS*. 1998.

[PM17] D. L. Poole and A. K. Mackworth. *Artificial intelligenceL foundations computational agents 2nd edition*. Cambridge University Press, 2017.

[PM18] J. Pearl and D. Mackenzie. *The book of why: the new science of cause and effect*. 2018.

[PMB19] J. Pérez, J. Marinkovic, and P. Barcelo. "On the Turing Completeness of Modern Neural Network Architectures". In: *ICLR*. 2019.

[Pog+17] T. Poggio, H. Mhaskar, L. Rosasco, B. Miranda, and Q. Liao. "Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review". en. In: *Int. J. Autom. Comput.* (2017), pp. 1–17.

[PP+20] M. Papadatou-Pastou, E. Ntolka, J. Schmitz, M. Martin, M. R. Munafò, S. Ocklenburg, and S. Paracchini. "Human handedness: A meta-analysis". en. In: *Psychol. Bull.* 146.6 (2020), pp. 481–524.

[PPS18] T. Pierrot, N. Perrin, and O. Sigaud. "First-order and second-order variants of the gradient descent in a unified framework". In: (2018). arXiv: 1810 . 08102 [cs.LG].

[Pre21] K. Pretz. "Stop Calling Everything AI, Machine-Learning Pioneer Says". In: *IEEE Spectrum* (2021).

[Pri+23] I. Price, A. Sanchez-Gonzalez, F. Alet, T. Ewalds, A. El-Kadi, J. Stott, S. Mohamed, P. Battaglia, R. Lam, and M. Willson. "GenCast: Diffusion-based ensemble forecasting for medium-range weather". In: (Dec. 2023). arXiv: 2312 . 15796 [cs.LG].

[PSM14a] J. Pennington, R. Socher, and C. Manning. "GloVe: Global vectors for word representation". In: *EMNLP*. 2014, pp. 1532–1543.

[PSM14b] J. Pennington, R. Socher, and C. Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[PSW15] N. G. Polson, J. G. Scott, and B. T. Willard. "Proximal Algorithms in Statistics and Machine Learning". en. In: *Stat. Sci.* 30.4 (2015), pp. 559–581.

[QC+06] J. Quiñonero-Candela, C. E. Rasmussen, F. Sinz, O. Bousquet, and B. Schölkopf. "Evaluating Predictive Uncertainty Challenge". In: *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2006, pp. 1–27.

[Qia+19] Q. Qian, L. Shang, B. Sun, J. Hu, H. Li, and R. Jin. "SoftTriple Loss: Deep Metric Learning Without Triplet Sampling". In: *ICCV*. 2019.

[Qiu+18] J. Qiu, Y. Dong, H. Ma, J. Li, K. Wang, and J. Tang. "Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 2018, pp. 459–467.

[Qiu+19a] J. Qiu, Y. Dong, H. Ma, J. Li, C. Wang, K. Wang, and J. Tang. "NetSMF: Large-Scale Network Embedding as Sparse Matrix Factorization". In: *The World Wide Web Conference*. WWW '19. Association for Computing Machinery, 2019, 1509–1520.

[Qiu+19b] J. Qiu, H. Ma, O. Levy, S. W. Yih, S. Wang, and J. Tang. "Blockwise Self-Attention for Long Document Understanding". In: *CoRR* abs/1911.02972 (2019). arXiv: 1911.02972.

[Qui86] J. R. Quinlan. "Induction of decision trees". In: *Machine Learning* 1 (1986), pp. 81–106.

[Qui93] J. R. Quinlan. *C4.5 Programs for Machine Learning*. Morgan Kauffman, 1993.

[Rad+] A. Radford et al. *Learning transferable visual models from natural language supervision*. Tech. rep. OpenAI.

[Rad+18] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. *Improving Language Understanding by Generative Pre-Training*. Tech. rep. OpenAI, 2018.

[Rad+19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. *Language Models are Unsupervised Multitask Learners*. Tech. rep. OpenAI, 2019.

[Raf+20] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *JMLR* (2020).

[Rag+17] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein. "On the Expressive Power of Deep Neural Networks". In: *ICML*. 2017.

[Rag+19] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. "Transfusion: Understanding transfer learning for medical imaging". In: *NIPS*. 2019, pp. 3347–3357.

[Rag+21] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy. "Do Vi-

sion Transformers See Like Convolutional Neural Networks?" In: *NIPS*. 2021.

[Raj+16] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. "SQuAD: 100,000+ Questions for Machine Comprehension of Text". In: *EMNLP*. 2016.

[Raj+18] A. Rajkomar et al. "Scalable and accurate deep learning with electronic health records". en. In: *NPJ Digit Med* 1 (2018), p. 18.

[Rat+09] M. Rattray, O. Stegle, K. Sharp, and J. Winn. "Inference algorithms and learning theory for Bayesian sparse factor analysis". In: *Proc. Intl. Workshop on Statistical-Mechanical Informatics*. 2009.

[RB93] M. Riedmiller and H. Braun. "A direct adaptive method for faster backpropagation learning: The RPROP algorithm". In: *ICNN*. IEEE. 1993, pp. 586–591.

[RBV17] S.-A. Rebuffi, H. Bilen, and A. Vedaldi. "Learning multiple visual domains with residual adapters". In: *NIPS*. 2017.

[RBV18] S.-A. Rebuffi, H. Bilen, and A. Vedaldi. "Efficient parametrization of multi-domain deep neural networks". In: *CVPR*. 2018.

[RC04] C. Robert and G. Casella. *Monte Carlo Statisical Methods*. 2nd edition. Springer, 2004.

[Rec+19] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. "Do Image Net Classifiers Generalize to Image Net?" In: *ICML*. 2019.

[Red+16] J Redmon, S Divvala, R Girshick, and A Farhadi. "You Only Look Once: Unified, Real-Time Object Detection". In: *CVPR*. 2016, pp. 779–788.

[Ren+09] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. "BPR: Bayesian Personalized Ranking from Implicit Feedback". In: *UAI*. 2009.

[Ren12] S. Rendle. "Factorization Machines with libFM". In: *ACM Trans. Intell. Syst. Technol.* 3.3 (2012), pp. 1–22.

[Ren19] Z. Ren. *List of papers on self-supervised learning*. 2019.

[Res+11] D. Reshef, Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti. "Detecting Novel Associations in Large Data Sets". In: *Science* 334 (2011), pp. 1518–1524.

[Res+16] Y. A. Reshef, D. N. Reshef, H. K. Finucane, P. C. Sabeti, and M. Mitzenmacher. "Measuring Dependence Powerfully and Equitably". In: *J. Mach. Learn. Res.* 17.211 (2016), pp. 1–63.

[RF17] J. Redmon and A. Farhadi. "YOLO9000: Better, Faster, Stronger". In: *CVPR*. 2017.

[RFB15] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *MICCAI (Intl. Conf. on Medical Image Computing and Computer Assisted Interventions)*. 2015.

[RG11] A. Rodriguez and K. Ghosh. *Modeling relational data through nested partition models*. Tech. rep. UC Santa Cruz, 2011.

[RHS05] C. Rosenberg, M. Hebert, and H. Schneiderman. "Semi-Supervised Self-Training of Object Detection Models". In: *Proceedings of the Seventh IEEE Workshops on Application of Computer Vision (WACV/MOTION'05)-Volume 1-Volume 01*. 2005, pp. 29–36.

[RHW86] D. Rumelhart, G. Hinton, and R. Williams. "Learning internal representations by error propagation". In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Ed. by D. Rumelhart, J. McClelland, and the PDD Research Group. MIT Press, 1986.

[Ric95] J. Rice. *Mathematical statistics and data analysis*. 2nd edition. Duxbury, 1995.

[Rif+11] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. "Contractive Auto-Encoders: Explicit Invariance During Feature Extraction". In: *ICML*. 2011.

[Ris+08] I. Rish, G. Grabarnik, G. Cecchi, F. Pereira, and G. Gordon. "Closed-form supervised dimensionality reduction with generalized linear models". In: *ICML*. 2008.

[RKK18] S. J. Reddi, S. Kale, and S. Kumar. "On the Convergence of Adam and Beyond". In: *ICLR*. 2018.

[RM01] N. Roy and A. McCallum. "Toward optimal active learning through Monte Carlo estimation of error reduction". In: *ICML*. 2001.

[RMC09] H. Rue, S. Martino, and N. Chopin. "Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations". In: *J. of Royal Stat. Soc. Series B* 71 (2009), pp. 319–392.

[RML22] S. Ramasinghe, L. Macdonald, and S. Lucey. "On the frequency-bias of coordinate-MLPs". In: *NIPS*. 2022.

[RMW14] D. J. Rezende, S. Mohamed, and D. Wierstra. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: *ICML*. Ed. by E. P. Xing and T. Jebara. Vol. 32. Proceedings of Machine Learning Research. PMLR, 2014, pp. 1278–1286.

[RN10] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. 3rd edition. Prentice Hall, 2010.

[Roo+21] F. de Roos, C. Jidling, A. Wills, T. Schön, and P. Hennig. "A Probabilistically Motivated Learning Rate Adaptation for Stochastic Optimization". In: (2021). arXiv: 2102.10880 [cs.LG].

[Ros58] F. Rosenblatt. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain". In: *Psychological Review* 65.6 (1958), pp. 386–408.

[Ros98] K. Rose. "Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems". In: *Proc. IEEE* 80 (1998), pp. 2210–2239.

[Rot+18] W. Roth, R. Peharz, S. Tschiatschek, and F. Pernkopf. "Hybrid generative-discriminative training of Gaussian mixture models". In: *Pattern Recognition Letters* 112 (2018), pp. 131–137.

[Rot+20] K. Roth, T. Milbich, S. Sinha, P. Gupta, B. Ommer, and J. P. Cohen. "Revisiting Training Strategies and Generalization Performance in Deep Metric Learning". In: *ICML*. 2020.

[Rou+09] J. Rouder, P. Speckman, D. Sun, and R. Morey. "Bayesian t tests for accepting and rejecting the null hypothesis". In: *Psychonomic Bulletin & Review* 16.2 (2009), pp. 225–237.

[Row97] S. Roweis. "EM algorithms for PCA and SPCA". In: *NIPS*. 1997.

[Roy+20] A. Roy, M. Saffar, A. Vaswani, and D. Grangier. "Efficient Content-Based Sparse Attention with Routing Transformers". In: *CoRR* abs/2003.05997 (2020). arXiv: 2003.05997.

[Roz+19] B. Rozemberczki, R. Davies, R. Sarkar, and C. Sutton. "GEMSEC: Graph Embedding with Self Clustering". In: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM '19. Association for Computing Machinery, 2019, 65–72.

[RP99] M. Riesenhuber and T. Poggio. "Hierarchical Models of Object Recognition in Cortex". In: *Nature Neuroscience* 2 (1999), pp. 1019–1025.

[RR08] A. Rahimi and B. Recht. "Random Features for Large-Scale Kernel Machines". In: *NIPS*. Curran Associates, Inc., 2008, pp. 1177–1184.

[RR09] A. Rahimi and B. Recht. "Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning". In: *NIPS*. Curran Associates, Inc., 2009, pp. 1313–1320.

[RS00] S. T. Roweis and L. K. Saul. "Nonlinear dimensionality reduction by locally linear embedding". en. In: *Science* 290.5500 (2000), pp. 2323–2326.

[RT82] D. B. Rubin and D. T. Thayer. "EM algorithms for ML factor analysis". In: *Psychometrika* 47.1 (1982), pp. 69–76.

[Rub84] D. B. Rubin. "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician". In: *Ann. Stat.* 12.4 (1984), pp. 1151–1172.

[Rup88] D Ruppert. *Efficient Estimations from a Slowly Convergent Robbins-Monro Process*. Tech. rep. 1988.

[Rus+15] O. Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *Intl. J. Computer Vision* (2015), pp. 1–42.

[Rus15] S. Russell. "Unifying Logic and Probability". In: *Commun. ACM* 58.7 (2015), pp. 88–97.

[Rus18] A. M. Rush. "The Annotated Transformer". In: *Proceedings of ACL Workshop on Open Source Software for NLP*. 2018.

[Rus19] S. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. en. Kindle. Viking, 2019.

[RW06] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[RZL17] P. Ramachandran, B. Zoph, and Q. V. Le. "Searching for Activation Functions". In: (2017). arXiv: 1710.05941 [cs.NE].

[SA93] P Sinha and E Adelson. "Recovering reflectance and illumination in a world of painted polyhedra". In: *ICCV*. 1993, pp. 156–163.

[Sab21] W. Saba. "Machine Learning Won't Solve Natural Language Understanding". In: (2021).

[Sal+16] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. "Improved Techniques for Training GANs". In: (2016). arXiv: 1606.03498 [cs.LG].

[SAM04] D. J. Spiegelhalter, K. R. Abrams, and J. P. Myles. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, 2004.

[San+18a] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. "Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation". In: (2018). arXiv: 1801.04381 [cs.CV].

[San+18b] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry. "How Does Batch Normal-

ization Help Optimization? (No, It Is Not About Internal Covariate Shift)". In: *NIPS*. 2018.

[San96] R. Santos. "Equivalence of regularization and truncated iteration for general ill-posed problems". In: *Linear Algebra and its Applications* 236.15 (1996), pp. 25–33.

[Sar11] R. Sarkar. "Low distortion delaunay embedding of trees in hyperbolic plane". In: *International Symposium on Graph Drawing*. Springer. 2011, pp. 355–366.

[SAV20] E. Stevens, L. Antiga, and T. Viehmann. *Deep Learning with PyTorch*. Manning, 2020.

[SBB01] T. Sellke, M. J. Bayarri, and J. Berger. "Calibration of p Values for Testing Precise Null Hypotheses". In: *The American Statistician* 55.1 (2001), pp. 62–71.

[SBP17] Y Sun, P Babu, and D. P. Palomar. "Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning". In: *IEEE Trans. Signal Process.* 65.3 (2017), pp. 794–816.

[SBS20] K. Shi, D. Bieber, and C. Sutton. "Incremental sampling without replacement for sequence models". In: *ICML*. 2020.

[Sca+09] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. "The graph neural network model". In: *IEEE Transactions on Neural Networks* 20.1 (2009), pp. 61–80.

[Sca+17] S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini. "Group Sparse Regularization for Deep Neural Networks". In: *Neurocomputing* 241 (2017).

[Sch+00] B Scholkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. "New support vector algorithms". en. In: *Neural Comput.* 12.5 (2000), pp. 1207–1245.

[Sch19] B. Schölkopf. "Causality for Machine Learning". In: (2019). arXiv: 1911.10500 [cs.LG].

[Sch78] G. Schwarz. "Estimating the dimension of a model". In: *Annals of Statistics* 6.2 (1978), pp. 461–464.

[Sch90] R. E. Schapire. "The strength of weak learnability". In: *Mach. Learn.* 5.2 (1990), pp. 197–227.

[Sco79] D. Scott. "On optimal and data-based histograms". In: *Biometrika* 66.3 (1979), pp. 605–610.

[Scu10] D Sculley. "Web-scale k-means clustering". In: *WWW*. WWW '10. Association for Computing Machinery, 2010, pp. 1177–1178.

[Scu65] H. Scudder. "Probability of error of some adaptive pattern-recognition machines". In: *IEEE Transactions on Information Theory* 11.3 (1965), pp. 363–371.

[Sed+15] S. Sedhain, A. K. Menon, S. Sanner, and L. Xie. "AutoRec: Autoencoders Meet Collaborative Filtering". In: *WWW*. WWW '15 Companion. Association for Computing Machinery, 2015, pp. 111–112.

[Sej18] T. J. Sejnowski. *The Deep Learning Revolution*. en. Kindle. The MIT Press, 2018.

[Set12] B. Settles. "Active learning". In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6 (2012), 1–114.

[SF12] R. Schapire and Y. Freund. *Boosting: Foundations and Algorithms*. MIT Press, 2012.

[SGJ11] D. Sontag, A. Globerson, and T. Jaakkola. "Introduction to Dual Decomposition for Inference". In: *Optimization for Machine Learning*. Ed. by S. Sra, S. Nowozin, and S. J. Wright. MIT Press, 2011.

[Sha+06] P. Shafto, C. Kemp, V. Mansinghka, M. Gordon, and J. B. Tenenbaum. "Learning cross-cutting systems of categories". In: *Cognitive Science Conference*. 2006.

[Sha+17] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer". In: *ICLR*. 2017.

[Sha88] T. Shallice. *From Neuropsychology to Mental Structure*. 1988.

[SHB16] R. Sennrich, B. Haddow, and A. Birch. "Neural Machine Translation of Rare Words with Subword Units". In: *Proc. ACL*. 2016.

[She+18] Z. Shen, M. Zhang, S. Yi, J. Yan, and H. Zhao. "Factorized Attention: Self-Attention with Linear Complexities". In: *CoRR* abs/1812.01243 (2018). arXiv: 1812.01243.

[She94] J. R. Shewchuk. *An introduction to the conjugate gradient method without the agonizing pain*. Tech. rep. CMU, 1994.

[SHF15] R. Steorts, R. Hall, and S. Fienberg. "A Bayesian Approach to Graphical Record Linkage and De-duplication". In: *JASA* (2015).

[Shu+13] D. I. Shuman, S. K. Narang, P Frossard, A Ortega, and P Vandergheynst. "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains". In: *IEEE Signal Process. Mag.* 30.3 (2013), pp. 83–98.

[Sin+20] S. Sinha, H. Zhang, A. Goyal, Y. Bengio, H. Larochelle, and A. Odena. "Small-GAN: Speeding up GAN Training using Core-Sets". In: *ICML*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 9005–9015.

[Sit+20] V. Sitzmann, J. N. P. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein. “Implicit Neural Representations with Periodic Activation Functions”. In: *NIPS*. https://www.vincentsitzmann.com/siren/. June 2020.

[SIV17] C. Szegedy, S. Ioffe, and V. Vanhoucke. “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning”. In: *AAAI*. 2017.

[SJ03] N. Srebro and T. Jaakkola. “Weighted low-rank approximations”. In: *ICML*. 2003.

[SJT16] M. Sajjadi, M. Javanmardi, and T. Tasdizen. “Regularization with stochastic transformations and perturbations for deep semi-supervised learning”. In: *Advances in neural information processing systems*. 2016, pp. 1163–1171.

[SK20] S. Singh and S. Krishnan. “Filter Response Normalization Layer: Eliminating Batch Dependence in the Training of Deep Neural Networks”. In: *CVPR*. 2020.

[SKP15] F. Schroff, D. Kalenichenko, and J. Philbin. “FaceNet: A Unified Embedding for Face Recognition and Clustering”. In: *CVPR*. 2015.

[SKT14] A. Szlam, Y. Kluger, and M. Tygert. “An implementation of a randomized algorithm for principal component analysis”. In: (2014). arXiv: 1412.3510 [stat.CO].

[SKTF18] H. Shao, A. Kumar, and P Thomas Fletcher. “The Riemannian Geometry of Deep Generative Models”. In: *CVPR*. 2018, pp. 315–323.

[SL18] S. L. Smith and Q. V. Le. “A Bayesian Perspective on Generalization and Stochastic Gradient Descent”. In: *ICLR*. 2018.

[SL+19] B. Sanchez-Lengeling, J. N. Wei, B. K. Lee, R. C. Gerkin, A. Aspuru-Guzik, and A. B. Wiltschko. “Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules”. In: (2019). arXiv: 1910.10685 [stat.ML].

[SL90] D. J. Spiegelhalter and S. L. Lauritzen. “Sequential updating of conditional probabilities on directed graphical structures”. In: *Networks* 20 (1990).

[SLRB17] M. Schmidt, N. Le Roux, and F. Bach. “Minimizing finite sums with the stochastic average gradient”. In: *Mathematical Programming* 162.1-2 (2017), pp. 83–112.

[SM00] J. Shi and J. Malik. “Normalized Cuts and Image Segmentation”. In: *IEEE PAMI* (2000).

[SM08] R. Salakhutdinov and A. Mnih. “Probabilistic Matrix Factorization”. In: *NIPS*. Vol. 20. 2008.

[SMG14] A. M. Saxe, J. L. McClelland, and S. Ganguli. “Exact solutions to the nonlinear dynamics of learning in deep linear neural networks”. In: *ICLR*. 2014.

[SMH07] R. Salakhutdinov, A. Mnih, and G. Hinton. “Restricted Boltzmann machines for collaborative filtering”. In: *ICML*. ICML ’07. Association for Computing Machinery, 2007, pp. 791–798.

[Smi18] L. Smith. “A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay”. In: (2018). arXiv: 1803.09820.

[Smi+21] S. L. Smith, B. Dherin, D. Barrett, and S. De. “On the Origin of Implicit Regularization in Stochastic Gradient Descent”. In: *ICLR*. 2021.

[SMM03] Q. Sheng, Y. Moreau, and B. D. Moor. “Biclustering Microarray data by Gibbs sampling”. In: *Bioinformatics* 19 (2003), pp. ii196–ii205.

[SNM16] M. Suzuki, K. Nakayama, and Y. Matsuo. “Joint Multimodal Learning with Deep Generative Models”. In: (2016). arXiv: 1611.01891 [stat.ML].

[Soh16] K. Sohn. “Improved Deep Metric Learning with Multi-class N-pair Loss Objective”. In: *NIPS*. Curran Associates, Inc., 2016, pp. 1857–1865.

[Soh+20] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel. “FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence”. In: (2020). arXiv: 2001.07685 [cs.LG].

[SP97] M Schuster and K. K. Paliwal. “Bidirectional recurrent neural networks”. In: *IEEE. Trans on Signal Processing* 45.11 (1997), pp. 2673–2681.

[Spe11] T. Speed. “A correlation for the 21st century”. In: *Science* 334 (2011), pp. 1502–1503.

[SR15] T. Saito and M. Rehmsmeier. “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets”. en. In: *PLoS One* 10.3 (2015), e0118432.

[SRG03] R. Salakhutdinov, S. T. Roweis, and Z. Ghahramani. “Optimization with EM and Expectation-Conjugate-Gradient”. In: *ICML*. 2003.

[Sri+14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *JMLR* (2014).

[SS01] B. Schlkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines,*

Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning). en. 1st edition. The MIT Press, 2001.

[SS02] B. Schoelkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[SS05] J. Schaefer and K. Strimmer. "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics". In: *Statist. Appl. Genet. Mol. Biol* 4.32 (2005).

[SS19] S. Serrano and N. A. Smith. "Is Attention Interpretable?" In: *Proc. ACL*. 2019.

[SS95] H. T. Siegelmann and E. D. Sontag. "On the Computational Power of Neural Nets". In: *J. Comput. System Sci.* 50.1 (1995), pp. 132–150.

[SSM98] B. Schoelkopf, A. Smola, and K.-R. Mueller. "Nonlinear component analysis as a kernel Eigenvalue problem". In: *Neural Computation* 10 (5 1998), pp. 1299 – 1319.

[Sta+06] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. "BioGRID: a general repository for interaction datasets". In: *Nucleic acids research* 34.suppl_1 (2006), pp. D535–D539.

[Ste56] C. Stein. "Inadmissibility of the usual estimator for the mean of a multivariate distribution". In: *Proc. 3rd Berkeley Symposium on Mathematical Statistics and Probability* (1956), 197–206.

[Str09] G. Strang. *Introduction to linear algebra*. 4th edition. SIAM Press, 2009.

[Sug+19] A. S. Suggala, K. Bhatia, P. Ravikumar, and P. Jain. "Adaptive Hard Thresholding for Near-optimal Consistent Robust Regression". In: *Proceedings of the Annual Conference On Learning Theory (COLT)*. 2019, pp. 2892–2897.

[Sun+09] L. Sun, S. Ji, S. Yu, and J. Ye. "On the Equivalence Between Canonical Correlation Analysis and Orthonormalized Partial Least Squares". In: *IJCAI*. 2009.

[Sun+19a] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. "VideoBERT: A Joint Model for Video and Language Representation Learning". In: *ICCV*. 2019.

[Sun+19b] S. Sun, Z. Cao, H. Zhu, and J. Zhao. "A Survey of Optimization Methods from a Machine Learning Perspective". In: (2019). arXiv: 1906.06821 [cs.LG].

[SVL14] I. Sutskever, O. Vinyals, and Q. V. V. Le. "Sequence to Sequence Learning with Neural Networks". In: *NIPS*. 2014.

[SVZ14] K. Simonyan, A. Vedaldi, and A. Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: *ICLR*. 2014.

[SW87] M. Shewry and H. Wynn. "Maximum entropy sampling". In: *J. Applied Statistics* 14 (1987), 165–170.

[SWY75] G Salton, A Wong, and C. S. Yang. "A vector space model for automatic indexing". In: *Commun. ACM* 18.11 (1975), pp. 613–620.

[Sze+15a] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going Deeper with Convolutions". In: *CVPR*. 2015.

[Sze+15b] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. "Rethinking the Inception Architecture for Computer Vision". In: (2015). arXiv: 1512.00567 [cs.CV].

[Tal07] N. Taleb. *The Black Swan: The Impact of the Highly Improbable*. Random House, 2007.

[Tan+15] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. "Line: Large-scale information network embedding". In: *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2015, pp. 1067–1077.

[Tan+18] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. "A Survey on Deep Transfer Learning". In: *ICANN*. 2018.

[Tan+20] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng. "Fourier features let networks learn high frequency functions in low dimensional domains". In: *NIPS*. June 2020.

[TAS18] M. Teye, H. Azizpour, and K. Smith. "Bayesian Uncertainty Estimation for Batch Normalized Deep Networks". In: *ICML*. 2018.

[Tay+20a] Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, and D. Metzler. "Long Range Arena: A Benchmark for efficient Transformers". In: *CoRR* (2020).

[Tay+20b] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler. "Efficient Transformers: A Survey". In: (2020). arXiv: 2009.06732 [cs.LG].

[TB97] L. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, 1997.

[TB99] M. Tipping and C. Bishop. "Probabilistic principal component analysis". In: *J.*

of Royal Stat. Soc. Series B 21.3 (1999), pp. 611–622.

[TDP19] I. Tenney, D. Das, and E. Pavlick. "BERT Rediscovers the Classical NLP Pipeline". In: *Proc. ACL*. 2019.

[TF03] M. Tipping and A. Faul. "Fast marginal likelihood maximisation for sparse Bayesian models". In: *AI/Stats*. 2003.

[Tho16] M. Thoma. "Creativity in Machine Learning". In: (2016). arXiv: 1601.03642 [cs.CV].

[Tho17] R. Thomas. *Computational Linear Algebra for Coders*. 2017.

[Tib96] R. Tibshirani. "Regression shrinkage and selection via the lasso". In: *J. Royal. Statist. Soc B* 58.1 (1996), pp. 267–288.

[Tip01] M. Tipping. "Sparse Bayesian learning and the relevance vector machine". In: *JMLR* 1 (2001), pp. 211–244.

[Tip98] M. Tipping. "Probabilistic visualization of high-dimensional binary data". In: *NIPS*. 1998.

[Tit16] M. Titsias. "One-vs-Each Approximation to Softmax for Scalable Estimation of Probabilities". In: *NIPS*. 2016, pp. 4161–4169.

[TK86] L. Tierney and J. Kadane. "Accurate approximations for posterior moments and marginal densities". In: *JASA* 81.393 (1986).

[TL21] M. Tan and Q. V. Le. "EfficientNetV2: Smaller Models and Faster Training". In: (2021). arXiv: 2104.00298 [cs.CV].

[TM15] D. Trafimow and M. Marks. "Editorial". In: *Basic Appl. Soc. Psych.* 37.1 (2015), pp. 1–2.

[TMP20] A. Tsitsulin, M. Munkhoeva, and B. Perozzi. "Just SLaQ When You Approximate: Accurate Spectral Distances for Web-Scale Graphs". In: *Proceedings of The Web Conference 2020*. WWW '20. 2020, 2697–2703.

[TOB16] L. Theis, A. van den Oord, and M. Bethge. "A note on the evaluation of generative models". In: *ICLR*. 2016.

[Tol+21] I. Tolstikhin et al. "MLP-Mixer: An all-MLP Architecture for Vision". In: (2021). arXiv: 2105.01601 [cs.CV].

[TP10] P. D. Turney and P. Pantel. "From Frequency to Meaning: Vector Space Models of Semantics". In: *JAIR* 37 (2010), pp. 141–188.

[TP97] S. Thrun and L. Pratt, eds. *Learning to learn*. Kluwer, 1997.

[TS92] D. G. Terrell and D. W. Scott. "Variable kernel density estimation". In: *Annals of Statistics* 20.3 (1992), 1236–1265.

[Tsi+18] A. Tsitsulin, D. Mottin, P. Karras, A. Bronstein, and E. Müller. "NetLSD: Hearing the Shape of a Graph". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '18. 2018, 2347–2356.

[TSL00] J. Tenenbaum, V. de Silva, and J. Langford. "A global geometric framework for nonlinear dimensionality reduction". In: *Science* 290.550 (2000), pp. 2319–2323.

[Tur13] M. Turk. "Over Twenty Years of Eigenfaces". In: *ACM Trans. Multimedia Comput. Commun. Appl.* 9.1s (2013), 45:1–45:5.

[TV17] A. Tarvainen and H. Valpola. "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results". In: *Advances in neural information processing systems*. 2017, pp. 1195–1204.

[TVW05] B. Turlach, W. Venables, and S. Wright. "Simultaneous Variable Selection". In: *Technometrics* 47.3 (2005), pp. 349–363.

[TW18] J. Tang and K. Wang. "Personalized Top-N Sequential Recommendation via Convolutional Sequence Embedding". In: *WSDM*. WSDM '18. Association for Computing Machinery, 2018, pp. 565–573.

[UB05] I. Ulusoy and C. M. Bishop. "Generative versus discriminative methods for object recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2005, pp. 258–265.

[Ude+16] M. Udell, C. Horn, R. Zadeh, and S. Boyd. "Generalized Low Rank Models". In: *Foundations and Trends in Machine Learning* 9.1 (2016), pp. 1–118.

[Uly+16] D. Ulyanov, V. Lebedev, Andrea, and V. Lempitsky. "Texture Networks: Feed-forward Synthesis of Textures and Stylized Images". In: *ICML*. 2016, pp. 1349–1357.

[Uur+17] V. Uurtio, J. M. Monteiro, J. Kandola, J. Shawe-Taylor, D. Fernandez-Reyes, and J. Rousu. "A Tutorial on Canonical Correlation Methods". In: *ACM Computing Surveys* (2017).

[UVL16] D. Ulyanov, A. Vedaldi, and V. Lempitsky. "Instance Normalization: The Missing Ingredient for Fast Stylization". In: (2016). arXiv: 1607.08022 [cs.CV].

[Van06] L. Vandenberghe. *Applied Numerical Computing: Lecture notes*. 2006.

[Van14] J. VanderPlas. *Frequentism and Bayesianism III: Confidence, Credibility, and why Frequentism and Science do not Mix*. Blog post. 2014.

[Van18] J. Vanschoren. "Meta-Learning: A Survey". In: (2018). arXiv: 1810 . 03548 [cs.LG].

[Vap98] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[Vas+17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. "Attention Is All You Need". In: *NIPS*. 2017.

[Vas+19] S. Vaswani, A. Mishkin, I. Laradji, M. Schmidt, G. Gidel, and S. Lacoste-Julien. "Painless Stochastic Gradient: Interpolation, Line-Search, and Convergence Rates". In: *NIPS*. Curran Associates, Inc., 2019, pp. 3727–3740.

[VD99] S. Vaithyanathan and B. Dom. "Model Selection in Unsupervised Learning With Applications To Document Clustering". In: *ICML*. 1999.

[VEB09] N. Vinh, J. Epps, and J. Bailey. "Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary?" In: *ICML*. 2009.

[Vel+18] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. "Graph attention networks". In: *ICLR*. 2018.

[Vel+19] P. Veličković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm. "Deep Graph Infomax". In: *International Conference on Learning Representations*. 2019.

[VGG17] A. Vehtari, A. Gelman, and J. Gabry. "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". In: *Stat. Comput.* 27.5 (2017), pp. 1413–1432.

[VGS97] V. Vapnik, S. Golowich, and A. Smola. "Support vector method for function approximation, regression estimation, and signal processing". In: *NIPS*. 1997.

[Vig15] T. Vigen. *Spurious Correlations*. en. Gift edition. Hachette Books, 2015.

[Vij+18] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra. "Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models". In: *IJCAI*. 2018.

[Vin+10a] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion". In: *JMLR* 11 (2010), pp. 3371–3408.

[Vin+10b] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion". In: *Journal of machine learning research* 11.Dec (2010), pp. 3371–3408.

[Vin+16] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra. "Matching Networks for One Shot Learning". In: *NIPS*. 2016.

[Vir10] S. Virtanen. "Bayesian exponential family projections". MA thesis. Aalto University, 2010.

[Vis+10] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgward. "Graph Kernels". In: *JMLR* 11 (2010), pp. 1201–1242.

[Vo+15] B.-N. Vo, M. Mallick, Y. Bar-Shalom, S. Coraluppi, R. Osborne, R. Mahler, B. t Vo, and J. Webster. *Multitarget tracking*. John Wiley and Sons, 2015.

[Vor+17] E. Vorontsov, C. Trabelsi, S. Kadoury, and C. Pal. "On orthogonality and learning recurrent networks with long term dependencies". In: *ICML*. 2017.

[VT17] C. Vondrick and A. Torralba. "Generating the Future with Adversarial Transformers". In: *CVPR*. 2017.

[VV13] G. Valiant and P. Valiant. "Estimating the unseen: improved estimators for entropy and other properties". In: *NIPS*. 2013.

[Wah+22] O. Wahltinez, A. Cheung, R. Alcantara, D. Cheung, M. Daswani, A. Erlinger, M. Lee, P. Yawalkar, M. P. Brenner, and K. Murphy. "COVID-19 Open-Data: a global-scale, spatially granular meta-dataset for SARS-CoV-2". In: *Nature Scientific Data* 9.162 (2022).

[Wal05] C. S. Wallace. *Statistical and Inductive Inference by Minimum Message Length (Information Science and Statistics)*. en. 2005th ed. Springer, May 2005.

[Wal+20] M. Walmsley et al. "Galaxy Zoo: probabilistic morphology through Bayesian CNNs and active learning". In: *Monthly Notices Royal Astronomial Society* 491.2 (2020), pp. 1554–1574.

[Wal47] A. Wald. "An Essentially Complete Class of Admissible Decision Functions". en. In: *Ann. Math. Stat.* 18.4 (1947), pp. 549–555.

[Wan+15] J. Wang, W. Liu, S. Kumar, and S.-F. Chang. "Learning to Hash for Indexing Big Data - A Survey". In: *Proc. IEEE* (2015).

[Wan+17] Y. Wang et al. "Tacotron: Towards End-to-End Speech Synthesis". In: *Interspeech*. 2017.

[Wan+20a] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma. "Linformer: Self-Attention with Linear Complexity". In: *CoRR* abs/2006.04768 (2020). arXiv: 2006.04768.

[Wan+20b] Y. Wang, Q. Yao, J. Kwok, and L. M. Ni. "Generalizing from a Few Examples: A Survey on Few-Shot Learning". In: *ACM Computing Surveys* 1.1 (2020).

[Wan+21] R. Wang, M. Cheng, X. Chen, X. Tang, and C.-J. Hsieh. “Rethinking Architecture Selection in Differentiable NAS”. In: *ICLR*. 2021.

[Was04] L. Wasserman. *All of statistics. A concise course in statistical inference*. Springer, 2004.

[Wat10] S. Watanabe. “Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory”. In: *JMLR* 11 (2010), pp. 3571–3594.

[Wat13] S. Watanabe. “A Widely Applicable Bayesian Information Criterion”. In: *JMLR* 14 (2013), pp. 867–897.

[WCS08] M. Welling, C. Chemudugunta, and N. Sutter. “Deterministic Latent Variable Models and their Pitfalls”. In: *ICDM*. 2008.

[WCZ16] D. Wang, P. Cui, and W. Zhu. “Structural deep network embedding”. In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2016, pp. 1225–1234.

[Wei76] J. Weizenbaum. *Computer Power and Human Reason: From Judgment to Calculation*. en. 1st ed. W H Freeman & Co, 1976.

[Wen+16] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li. “Learning Structured Sparsity in Deep Neural Networks”. In: (2016). arXiv: 1608.03665 [cs.NE].

[Wen18] L. Weng. “Attention? Attention!” In: *lilianweng.github.io/lil-log* (2018).

[Wen19] L. Weng. “Generalized Language Models”. In: *lilianweng.github.io/lil-log* (2019).

[Wer74] P. Werbos. “Beyond regression: New Tools for Prediction and Analysis in the Behavioral Sciences”. PhD thesis. Harvard, 1974.

[Wer90] P. J. Werbos. “Backpropagation Through Time: What It Does and How to Do It”. In: *Proc. IEEE* 78.10 (1990), pp. 1550–1560.

[Wes03] M. West. “Bayesian Factor Regression Models in the "Large p, Small n" Paradigm”. In: *Bayesian Statistics 7* (2003).

[WF14] Z. Wang and N. de Freitas. “Theoretical Analysis of Bayesian Optimisation with Unknown Gaussian Process Hyper-Parameters”. In: (2014). arXiv: 1406.7758 [stat.ML].

[WF20] T. Wu and I. Fischer. “Phase Transitions for the Information Bottleneck in Representation Learning”. In: *ICLR*. 2020.

[WH18] Y. Wu and K. He. “Group Normalization”. In: *ECCV*. 2018.

[WH60] B. Widrow and M. E. Hoff. “Adaptive Switching Circuits”. In: *1960 IRE WESCON Convention Record, Part 4*. IRE, 1960, pp. 96–104.

[WI20] A. G. Wilson and P. Izmailov. “Bayesian Deep Learning and a Probabilistic Perspective of Generalization”. In: *NIPS*. 2020.

[Wil14] A. G. Wilson. “Covariance kernels for fast automatic pattern discovery and extrapolation with Gaussian processes”. PhD thesis. University of Cambridge, 2014.

[Wil20] C. K. I. Williams. “The Effect of Class Imbalance on Precision-Recall Curves”. In: *Neural Comput.* (2020).

[WL08] T. T. Wu and K. Lange. “Coordinate descent algorithms for lasso penalized regression”. In: *Ann. Appl. Stat* 2.1 (2008), pp. 224–244.

[WLL16] W. Wang, H. Lee, and K. Livescu. “Deep Variational Canonical Correlation Analysis”. In: *arXiv* (2016).

[WM00] D. R. Wilson and T. R. Martinez. “Reduction Techniques for Instance-Based Learning Algorithms”. In: *Mach. Learn.* 38.3 (2000), pp. 257–286.

[WNF09] S. Wright, R. Nowak, and M. Figueiredo. “Sparse reconstruction by separable approximation”. In: *IEEE Trans. on Signal Processing* 57.7 (2009), pp. 2479–2493.

[WNS19] C. White, W. Neiswanger, and Y. Savani. “BANANAS: Bayesian Optimization with Neural Architectures for Neural Architecture Search”. In: (2019). arXiv: 1910.11858 [cs.LG].

[Wol92] D. Wolpert. “Stacked Generalization”. In: *Neural Networks* 5.2 (1992), pp. 241–259.

[Wol96] D. Wolpert. “The lack of a priori distinctions between learning algorithms”. In: *Neural Computation* 8.7 (1996), pp. 1341–1390.

[WP19] S. Wiegreffe and Y. Pinter. “Attention is not not Explanation”. In: *EMNLP*. 2019.

[WRC08] J. Weston, F. Ratle, and R. Collobert. “Deep learning via semi-supervised embedding”. In: *Proceedings of the 25th international conference on Machine learning*. ACM. 2008, pp. 1168–1175.

[WS09] K. Weinberger and L. Saul. “Distance Metric Learning for Large Margin Classification”. In: *JMLR* 10 (2009), pp. 207–244.

[WSH16] L. Wu, C. Shen, and A. van den Hengel. “PersonNet: Person Re-identification with Deep Convolutional Neural Networks”. In: (2016). arXiv: 1601.07255 [cs.CV].

[WSL19] R. L. Wasserstein, A. L. Schirm, and N. A. Lazar. “Moving to a World Beyond

"p < 0.05"". In: *The American Statistician* 73.sup1 (2019), pp. 1–19.

[WSS04] K. Q. Weinberger, F. Sha, and L. K. Saul. "Learning a kernel matrix for nonlinear dimensionality reduction". In: *ICML*. 2004.

[WTN19] Y. Wu, G. Tucker, and O. Nachum. "The Laplacian in RL: Learning Representations with Efficient Approximations". In: *ICLR*. 2019.

[Wu+16] Y. Wu et al. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation". In: (2016). arXiv: 1609 . 08144 [cs.CL].

[Wu+19] Y. Wu, E. Winston, D. Kaushik, and Z. Lipton. "Domain Adaptation with Asymmetrically-Relaxed Distribution Alignment". In: *ICML*. 2019.

[WVJ16] M. Wattenberg, F. Viégas, and I. Johnson. "How to Use t-SNE Effectively". In: *Distill* 1.10 (2016).

[WW93] D. Wagner and F. Wagner. "Between min cut and graph bisection". In: *Proc. 18th Intl. Symp. on Math. Found. of Comp. Sci.* 1993, pp. 744–750.

[Xie+19] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le. "Unsupervised data augmentation for consistency training". In: *arXiv preprint arXiv:1904.12848* (2019).

[Xie+20] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. "Self-training with noisy student improves imagenet classification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 10687–10698.

[XJ96] L. Xu and M. I. Jordan. "On Convergence Properties of the EM Algorithm for Gaussian Mixtures". In: *Neural Computation* 8 (1996), pp. 129–151.

[XRV17] H. Xiao, K. Rasul, and R. Vollgraf. "Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms". In: (2017). arXiv: 1708.07747 [stat.ML].

[Xu+15] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". In: *ICML*. 2015.

[Yal+19] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan. "Billion-scale semi-supervised learning for image classification". In: *arXiv preprint arXiv:1905.00546* (2019).

[Yan+14] X. Yang, Y. Guo, Y. Liu, and H. Steck. "A Survey of Collaborative Filtering Based Social Recommender Systems". In: *Comput. Commun.* 41 (2014), pp. 1–10.

[YB19] C. Yadav and L. Bottou. "Cold Case: The Lost MNIST Digits". In: *arXiv* (2019).

[YCS16] Z. Yang, W. W. Cohen, and R. Salakhutdinov. "Revisiting semi-supervised learning with graph embeddings". In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*. JMLR. org. 2016, pp. 40–48.

[Yeu91] R. W. Yeung. "A new outlook on Shannon's information measures". In: *IEEE Trans. Inf. Theory* 37.3 (1991), pp. 466–474.

[YHJ09] D. Yan, L. Huang, and M. I. Jordan. "Fast approximate spectral clustering". In: *15th ACM Conf. on Knowledge Discovery and Data Mining*. 2009.

[Yin+19] P. Yin, J. Lyu, S. Zhang, S. Osher, Y. Qi, and J. Xin. "Understanding Straight-Through Estimator in Training Activation Quantized Neural Nets". In: *ICLR*. 2019.

[YK16] F. Yu and V. Koltun. "Multi-Scale Context Aggregation by Dilated Convolutions". In: *ICLR*. 2016.

[YL06] M. Yuan and Y. Lin. "Model Selection and Estimation in Regression with Grouped Variables". In: *J. Royal Statistical Society, Series B* 68.1 (2006), pp. 49–67.

[YL21] A. L. Yuille and C. Liu. "Deep Nets: What have they ever done for Vision?" In: *Intl. J. Computer Vision* 129 (2021), pp. 781–802.

[Yon19] E. Yong. "The Human Brain Project Hasn't Lived Up to Its Promise". In: *The Atlantic* (2019).

[Yos+15] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. "Understanding Neural Networks Through Deep Visualization". In: *ICML Workshop on Deep Learning*. 2015.

[Yu+06] S. Yu, K. Yu, V. Tresp, K. H-P., and M. Wu. "Supervised probabilistic principal component analysis". In: *KDD*. 2006.

[Yu+16] F. X. X. Yu, A. T. Suresh, K. M. Choromanski, D. N. Holtmann-Rice, and S. Kumar. "Orthogonal Random Features". In: *NIPS*. Curran Associates, Inc., 2016, pp. 1975–1983.

[YWG12] S. E. Yuksel, J. N. Wilson, and P. D. Gader. "Twenty Years of Mixture of Experts". In: *IEEE Trans. on neural networks and learning systems* (2012).

[Zah+18] M. Zaheer, S. Reddi, D. Sachan, S. Kale, and S. Kumar. "Adaptive Methods for Nonconvex Optimization". In: *NIPS*.

Curran Associates, Inc., 2018, pp. 9815–9825.

[Zah+20] M. Zaheer et al. “Big Bird: Transformers for Longer Sequences”. In: *CoRR* abs/2007.14062 (2020). arXiv: 2007 . 14062.

[Zei12] M. D. Zeiler. “ADADELTA: An Adaptive Learning Rate Method”. In: (2012). arXiv: 1212.5701 [cs.LG].

[Zel76] A. Zellner. “Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms”. In: *JASA* 71.354 (1976), pp. 400–405.

[ZG02] X. Zhu and Z. Ghahramani. *Learning from labeled and unlabeled data with label propagation*. Tech. rep. CALD tech report CMU-CALD-02-107. CMU, 2002.

[ZG06] M. Zhu and A. Ghodsi. “Automatic dimensionality selection from the scree plot via the use of profile likelihood”. In: *Computational Statistics & Data Analysis* 51 (2006), pp. 918–930.

[ZH05] H. Zou and T. Hastie. “Regularization and Variable Selection via the Elastic Net”. In: *J. of Royal Stat. Soc. Series B* 67.2 (2005), pp. 301–320.

[Zha+17a] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. “Understanding deep learning requires rethinking generalization”. In: *ICLR*. 2017.

[Zha+17b] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. “mixup: Beyond Empirical Risk Minimization”. In: *ICLR*. 2017.

[Zha+18] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu. “Object Detection with Deep Learning: A Review”. In: (2018). arXiv: 1807.05511 [cs.CV].

[Zha+19a] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu. “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization”. In: (2019). arXiv: 1912.08777 [cs.CL].

[Zha+19b] S. Zhang, L. Yao, A. Sun, and Y. Tay. “Deep Learning Based Recommender System: A Survey and New Perspectives”. In: *ACM Comput. Surv.* 52.1 (2019), pp. 1–38.

[Zha+20] A. Zhang, Z. Lipton, M. Li, and A. Smola. *Dive into deep learning*. 2020.

[Zho+04] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. “Learning with local and global consistency”. In: *Advances in neural information processing systems*. 2004, pp. 321–328.

[Zho+18] D. Zhou, Y. Tang, Z. Yang, Y. Cao, and Q. Gu. “On the Convergence of Adaptive Gradient Methods for Nonconvex Optimization”. In: (2018). arXiv: 1808.05671 [cs.LG].

[Zho+21] C. Zhou, X. Ma, P. Michel, and G. Neubig. “Examining and Combating Spurious Features under Distribution Shift”. In: *ICML*. 2021.

[ZHT06] H. Zou, T. Hastie, and R. Tibshirani. “Sparse principal component analysis”. In: *JCGS* 15.2 (2006), pp. 262–286.

[Zhu05] X. Zhu. “Semi-supervised learning with graphs”. PhD thesis. Carnegie Mellon University, 2005.

[Zhu+19] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. “A Comprehensive Survey on Transfer Learning”. In: (2019). arXiv: 1911.02685 [cs.LG].

[Zie+05] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. “Improving recommendation lists through topic diversification”. In: *WWW*. WWW ’05. Association for Computing Machinery, 2005, pp. 22–32.

[ZK16] S. Zagoruyko and N. Komodakis. “Wide Residual Networks”. In: *BMVC*. 2016.

[ZL05] Z.-H. Zhou and M. Li. “Tri-training: Exploiting unlabeled data using three classifiers”. In: *IEEE Transactions on knowledge and Data Engineering* 17.11 (2005), pp. 1529–1541.

[ZL17] B. Zoph and Q. V. Le. “Neural Architecture Search with Reinforcement Learning”. In: *ICLR*. 2017.

[ZLZ20] D. Zhang, Y. Li, and Z. Zhang. “Deep metric learning with spherical embedding”. In: *NIPS*. 2020.

[ZMY19] D. Zabihzadeh, R. Monsefi, and H. S. Yazdi. “Sparse Bayesian approach for metric learning in latent space”. In: *Knowledge-Based Systems* 178 (2019), pp. 11–24.

[ZRY05] P. Zhao, G. Rocha, and B. Yu. *Grouped and Hierarchical Model Selection through Composite Absolute Penalties*. Tech. rep. UC Berkeley, 2005.

[ZS14] H. Zen and A Senior. “Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis”. In: *ICASSP*. 2014, pp. 3844–3848.

[ZY08] J.-H. Zhao and P. L. H. Yu. “Fast ML Estimation for the Mixture of Factor Analyzers via an ECM Algorithm”. In: *IEEE. Trans. on Neural Networks* 19.11 (2008).