References

Abernethy, J., Bartlett, P. L., Rakhlin, A. & Tewari, A. (2008), "Optimal strategies and minimax lower bounds for online convex games," in *Proceedings of the nineteenth annual conference on computational learning theory*.

Ackerman, M. & Ben-David, S. (2008), "Measures of clustering quality: A working set of axioms for clustering," in *Proceedings of Neural Information Processing Systems* (NIPS), pp. 121–128.

Agarwal, S. & Roth, D. (2005), "Learnability of bipartite ranking functions," in *Proceedings of the 18th annual conference on learning theory*, pp. 16–31.

Agmon, S. (1954), "The relaxation method for linear inequalities," *Canadian Journal of Mathematics* **6**(3), 382–392.

Aizerman, M. A., Braverman, E. M. & Rozonoer, L. I. (1964), "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and Remote Control* **25**, 821–837.

Allwein, E. L., Schapire, R. & Singer, Y. (2000), "Reducing multiclass to binary: A unifying approach for margin classifiers," *Journal of Machine Learning Research* **1**, 113–141.

Alon, N., Ben-David, S., Cesa-Bianchi, N. & Haussler, D. (1997), "Scale-sensitive dimensions, uniform convergence, and learnability," *Journal of the ACM* **44**(4), 615–631.

Anthony, M. & Bartlet, P. (1999), *Neural Network Learning: Theoretical Foundations*, Cambridge University Press.

Baraniuk, R., Davenport, M., DeVore, R. & Wakin, M. (2008), "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation* **28**(3), 253–263.

Barber, D. (2012), *Bayesian reasoning and machine learning*, Cambridge University Press.

Bartlett, P., Bousquet, O. & Mendelson, S. (2005), "Local rademacher complexities," *Annals of Statistics* **33**(4), 1497–1537.

Bartlett, P. L. & Ben-David, S. (2002), "Hardness results for neural network approximation problems," *Theor. Comput. Sci.* **284**(1), 53–66.

Bartlett, P. L., Long, P. M. & Williamson, R. C. (1994), "Fat-shattering and the learnability of real-valued functions," in *Proceedings of the seventh annual conference on computational learning theory*, (ACM), pp. 299–310.

Downloaded from https://www.cambridge.org/core. Bibliothekssystem Universitaet Hamburg, on 21 Apr 2019 at 16:48:28, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781107298019.036

Bartlett, P. L. & Mendelson, S. (2001), "Rademacher and Gaussian complexities: Risk bounds and structural results," in *14th Annual Conference on Computational Learning Theory* (COLT) *2001*, Vol. 2111, Springer, Berlin, pp. 224–240.

Bartlett, P. L. & Mendelson, S. (2002), "Rademacher and Gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research* **3**, 463–482.

Ben-David, S., Cesa-Bianchi, N., Haussler, D. & Long, P. (1995), "Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions," *Journal of Computer and System Sciences* **50**, 74–86.

Ben-David, S., Eiron, N. & Long, P. (2003), "On the difficulty of approximately maximizing agreements," *Journal of Computer and System Sciences* **66**(3), 496–514.

Ben-David, S. & Litman, A. (1998), "Combinatorial variability of vapnik-chervonenkis classes with applications to sample compression schemes," *Discrete Applied Mathematics* **86**(1), 3–25.

Ben-David, S., Pal, D., & Shalev-Shwartz, S. (2009), "Agnostic online learning," in Conference on Learning Theory (COLT).

Ben-David, S. & Simon, H. (2001), "Efficient learning of linear perceptrons," *Advances in Neural Information Processing Systems*, pp. 189–195.

Bengio, Y. (2009), "Learning deep architectures for AI," *Foundations and Trends in Machine Learning* **2**(1), 1–127.

Bengio, Y. & LeCun, Y. (2007), "Scaling learning algorithms towards AI," *Large-Scale Kernel Machines* **34**.

Bertsekas, D. (1999), *Nonlinear programming*, Athena Scientific.

Beygelzimer, A., Langford, J. & Ravikumar, P. (2007), "Multiclass classification with filter trees," *Preprint, June* .

Birkhoff, G. (1946), "Three observations on linear algebra," *Revi. Univ. Nac. Tucuman, ser. A* **5**, 147–151.

Bishop, C. M. (2006), *Pattern recognition and machine learning*, Vol. 1, Springer: New York.

Blum, L., Shub, M. & Smale, S. (1989), "On a theory of computation and complexity over the real numbers: Np-completeness, recursive functions and universal machines," *Am. Math. Soc.* **21**(1), 1–46.

Blumer, A., Ehrenfeucht, A., Haussler, D. & Warmuth, M. K. (1987), "Occam's razor," *Information Processing Letters* **24**(6), 377–380.

Blumer, A., Ehrenfeucht, A., Haussler, D. & Warmuth, M. K. (1989), "Learnability and the Vapnik-Chervonenkis dimension," *Journal of the Association for Computing Machinery* **36**(4), 929–965.

Borwein, J. & Lewis, A. (2006), *Convex analysis and nonlinear optimization*, Springer.

Boser, B. E., Guyon, I. M. & Vapnik, V. N. (1992), "A training algorithm for optimal margin classifiers," in COLT, pp. 144–152.

Bottou, L. & Bousquet, O. (2008), "The tradeoffs of large scale learning," in NIPS, pp. 161–168.

Boucheron, S., Bousquet, O. & Lugosi, G. (2005), "Theory of classification: A survey of recent advances," *ESAIM: Probability and Statistics* **9**, 323–375.

Bousquet, O. (2002), Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms, PhD thesis, Ecole Polytechnique.

Bousquet, O. & Elisseeff, A. (2002), "Stability and generalization," *Journal of Machine Learning Research* **2**, 499–526.

Boyd, S. & Vandenberghe, L. (2004), *Convex optimization*, Cambridge University Press.

Breiman, L. (1996), Bias, variance, and arcing classifiers, Technical Report 460, Statistics Department, University of California at Berkeley.

Breiman, L. (2001), "Random forests," *Machine Learning* **45**(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984), *Classification and regression trees*, Wadsworth & Brooks.

Downloaded from https://www.cambridge.org/core. Bibliothekssystem Universitaet Hamburg, on 21 Apr 2019 at 16:48:28, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781107298019.036

Candès, E. (2008), "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique* **346**(9), 589–592.

Candes, E. J. (2006), "Compressive sampling," in *Proc. of the int. congress of math.*, Madrid, Spain.

Candes, E. & Tao, T. (2005), "Decoding by linear programming," *IEEE Trans. on Information Theory* **51**, 4203–4215.

Cesa-Bianchi, N. & Lugosi, G. (2006), *Prediction, learning, and games*, Cambridge University Press.

Chang, H. S., Weiss, Y. & Freeman, W. T. (2009), "Informative sensing," *arXiv preprint arXiv:0901.4275*.

Chapelle, O., Le, Q. & Smola, A. (2007), "Large margin optimization of ranking measures," in *NIPS workshop: Machine learning for Web search* (Machine Learning).

Collins, M. (2000), "Discriminative reranking for natural language parsing," in Machine Learning.

Collins, M. (2002), "Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms," in *Conference on Empirical Methods in Natural Language Processing*.

Collobert, R. & Weston, J. (2008), "A unified architecture for natural language processing: deep neural networks with multitask learning," in International Conference on Machine Learning (ICML).

Cortes, C. & Vapnik, V. (1995), "Support-vector networks," *Machine Learning* **20**(3), 273–297.

Cover, T. (1965), "Behavior of sequential predictors of binary sequences," *Trans. 4th Prague conf. information theory statistical decision functions, random processes*, pp. 263–272.

Cover, T. & Hart, P. (1967), "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions on* **13**(1), 21–27.

Crammer, K. & Singer, Y. (2001), "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research* **2**, 265–292.

Cristianini, N. & Shawe-Taylor, J. (2000), *An introduction to support vector machines*, Cambridge University Press.

Daniely, A., Sabato, S., Ben-David, S. & Shalev-Shwartz, S. (2011), "Multiclass learnability and the erm principle," in COLT.

Daniely, A., Sabato, S. & Shwartz, S. S. (2012), "Multiclass learning approaches: A theoretical comparison with implications," in NIPS.

Davis, G., Mallat, S. & Avellaneda, M. (1997), "Greedy adaptive approximation," *Journal of Constructive Approximation* **13**, 57–98.

Devroye, L. & Györfi, L. (1985), *Nonparametric density estimation: The L B1 S view*, Wiley.

Devroye, L., Györfi, L. & Lugosi, G. (1996), *A probabilistic theory of pattern recognition*, Springer.

Dietterich, T. G. & Bakiri, G. (1995), "Solving multiclass learning problems via error-correcting output codes," *Journal of Artificial Intelligence Research* **2**, 263–286.

Donoho, D. L. (2006), "Compressed sensing," *Information Theory, IEEE Transactions* **52**(4), 1289–1306.

Dudley, R., Gine, E. & Zinn, J. (1991), "Uniform and universal glivenko-cantelli classes," *Journal of Theoretical Probability* **4**(3), 485–510.

Dudley, R. M. (1987), "Universal Donsker classes and metric entropy," *Annals of Probability* **15**(4), 1306–1326.

Fisher, R. A. (1922), "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **222**, 309–368.

Downloaded from https://www.cambridge.org/core. Bibliothekssystem Universitaet Hamburg, on 21 Apr 2019 at 16:48:28, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781107298019.036

Floyd, S. (1989), "Space-bounded learning and the Vapnik-Chervonenkis dimension," in COLT, pp. 349–364.

Floyd, S. & Warmuth, M. (1995), "Sample compression, learnability, and the Vapnik-Chervonenkis dimension," *Machine Learning* **21**(3), 269–304.

Frank, M. & Wolfe, P. (1956), "An algorithm for quadratic programming," *Naval Res. Logist. Quart.* **3**, 95–110.

Freund, Y. & Schapire, R. (1995), "A decision-theoretic generalization of on-line learning and an application to boosting," in European Conference on Computational Learning Theory (EuroCOLT), Springer-Verlag, pp. 23–37.

Freund, Y. & Schapire, R. E. (1999), "Large margin classification using the perceptron algorithm," *Machine Learning* **37**(3), 277–296.

Garcia, J. & Koelling, R. (1996), "Relation of cue to consequence in avoidance learning," *Foundations of animal behavior: classic papers with commentaries* **4**, 374.

Gentile, C. (2003), "The robustness of the p-norm algorithms," *Machine Learning* **53**(3), 265–299.

Georghiades, A., Belhumeur, P. & Kriegman, D. (2001), "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intelligence* **23**(6), 643–660.

Gordon, G. (1999), "Regret bounds for prediction problems," in Conference on Learning Theory (COLT).

Gottlieb, L.-A., Kontorovich, L. & Krauthgamer, R. (2010), "Efficient classification for metric data," in *23rd conference on learning theory*, pp. 433–440.

Guyon, I. & Elisseeff, A. (2003), "An introduction to variable and feature selection," *Journal of Machine Learning Research, Special Issue on Variable and Feature Selection* **3**, 1157–1182.

Hadamard, J. (1902), "Sur les problèmes aux dérivées partielles et leur signification physique," *Princeton University Bulletin* **13**, 49–52.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The elements of statistical learning*, Springer.

Haussler, D. (1992), "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Information and Computation* **100**(1), 78–150.

Haussler, D. & Long, P. M. (1995), "A generalization of sauer's lemma," *Journal of Combinatorial Theory, Series A* **71**(2), 219–240.

Hazan, E., Agarwal, A. & Kale, S. (2007), "Logarithmic regret algorithms for online convex optimization," *Machine Learning* **69**(2–3), 169–192.

Hinton, G. E., Osindero, S. & Teh, Y.-W. (2006), "A fast learning algorithm for deep belief nets," *Neural Computation* **18**(7), 1527–1554.

Hiriart-Urruty, J.-B. & Lemaréchal, C. (1993), *Convex analysis and minimization algorithms*, Springer.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003), "A practical guide to support vector classification."

Hyafil, L. & Rivest, R. L. (1976), "Constructing optimal binary decision trees is NP-complete," *Information Processing Letters* **5**(1), 15–17.

Joachims, T. (2005), "A support vector method for multivariate performance measures," in *Proceedings of the international conference on machine learning* (ICML).

Kakade, S., Sridharan, K. & Tewari, A. (2008), "On the complexity of linear prediction: Risk bounds, margin bounds, and regularization," in NIPS.

Karp, R. M. (1972), *Reducibility among combinatorial problems*, Springer.

Kearns, M. & Mansour, Y. (1996), "On the boosting ability of top-down decision tree learning algorithms," in ACM Symposium on the Theory of Computing (STOC).

Kearns, M. & Ron, D. (1999), "Algorithmic stability and sanity-check bounds for leave-one-out cross-validation," *Neural Computation* **11**(6), 1427–1453.

Downloaded from https://www.cambridge.org/core. Bibliothekssystem Universitaet Hamburg, on 21 Apr 2019 at 16:48:28, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781107298019.036

Kearns, M. & Valiant, L. G. (1988), "Learning Boolean formulae or finite automata is as hard as factoring, Technical Report TR-14-88, Harvard University, Aiken Computation Laboratory.

Kearns, M. & Vazirani, U. (1994), *An Introduction to Computational Learning Theory*, MIT Press.

Kearns, M. J., Schapire, R. E. & Sellie, L. M. (1994), "Toward efficient agnostic learning," *Machine Learning* **17**, 115–141.

Kleinberg, J. (2003), "An impossibility theorem for clustering," NIPS, pp. 463–470.

Klivans, A. R. & Sherstov, A. A. (2006), Cryptographic hardness for learning intersections of halfspaces, in FOCS.

Koller, D. & Friedman, N. (2009), *Probabilistic graphical models: Principles and techniques*, MIT Press.

Koltchinskii, V. & Panchenko, D. (2000), "Rademacher processes and bounding the risk of function learning," in *High Dimensional Probability II*, Springer, pp. 443–457.

Kuhn, H. W. (1955), "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly* **2**(1–2), 83–97.

Kutin, S. & Niyogi, P. (2002), "Almost-everywhere algorithmic stability and generalization error," in *Proceedings of the 18th conference in uncertainty in artificial intelligence*, pp. 275–282.

Lafferty, J., McCallum, A. & Pereira, F. (2001), "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *International conference on machine learning*, pp. 282–289.

Langford, J. (2006), "Tutorial on practical prediction theory for classification," *Journal of machine learning research* **6**(1), 273.

Langford, J. & Shawe-Taylor, J. (2003), "PAC-Bayes & margins," in NIPS, pp. 423–430.

Le, Q. V., Ranzato, M.-A., Monga, R., Devin, M., Corrado, G., Chen, K., Dean, J. & Ng, A. Y. (2012), "Building high-level features using large scale unsupervised learning," in ICML.

Le Cun, L. (2004), "Large scale online learning," in *Advances in neural information processing systems 16: Proceedings of the 2003 conference*, Vol. 16, MIT Press, p. 217.

LeCun, Y. & Bengio, Y. (1995), "Convolutional networks for images, speech, and time series," in *The handbook of brain theory and neural networks*, The MIT Press.

Lee, H., Grosse, R., Ranganath, R. & Ng, A. (2009), "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in ICML.

Littlestone, N. (1988), "Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm," *Machine Learning* **2**, 285–318.

Littlestone, N. & Warmuth, M. (1986), Relating data compression and learnability. Unpublished manuscript.

Littlestone, N. & Warmuth, M. K. (1994), "The weighted majority algorithm," *Information and Computation* **108**, 212–261.

Livni, R., Shalev-Shwartz, S. & Shamir, O. (2013), "A provably efficient algorithm for training deep networks," *arXiv preprint arXiv:1304.7045* .

Livni, R. & Simon, P. (2013), "Honest compressions and their application to compression schemes," in COLT.

MacKay, D. J. (2003), *Information theory, inference and learning algorithms*, Cambridge University Press.

Mallat, S. & Zhang, Z. (1993), "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing* **41**, 3397–3415.

McAllester, D. A. (1998), "Some PAC-Bayesian theorems," in COLT.

McAllester, D. A. (1999), "PAC-Bayesian model averaging," in COLT, pp. 164–170.

McAllester, D. A. (2003), "Simplified PAC-Bayesian margin bounds," in COLT, pp. 203–215.

Downloaded from https://www.cambridge.org/core. Bibliothekssystem Universitaet Hamburg, on 21 Apr 2019 at 16:48:28, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781107298019.036

Minsky, M. & Papert, S. (1969), *Perceptrons: An introduction to computational geometry*, The MIT Press.

Mukherjee, S., Niyogi, P., Poggio, T. & Rifkin, R. (2006), "Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization," *Advances in Computational Mathematics* **25**(1–3), 161–193.

Murata, N. (1998), "A statistical study of on-line learning," *Online Learning and Neural Networks, Cambridge University Press*.

Murphy, K. P. (2012), *Machine learning: a probabilistic perspective*, The MIT Press.

Natarajan, B. (1995), "Sparse approximate solutions to linear systems," *SIAM J. Computing* **25**(2), 227–234.

Natarajan, B. K. (1989), "On learning sets and functions," *Mach. Learn.* **4**, 67–97.

Nemirovski, A., Juditsky, A., Lan, G. & Shapiro, A. (2009), "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on Optimization* **19**(4), 1574–1609.

Nemirovski, A. & Yudin, D. (1978), *Problem complexity and method efficiency in optimization*, Nauka, Moscow.

Nesterov, Y. (2005), Primal-dual subgradient methods for convex problems, Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL).

Nesterov, Y. & Nesterov, I. (2004), *Introductory lectures on convex optimization: A basic course*, Vol. 87, Springer, Netherlands.

Novikoff, A. B. J. (1962), "On convergence proofs on perceptrons," in *Proceedings of the symposium on the mathematical theory of automata*, Vol. XII, pp. 615–622.

Parberry, I. (1994), *Circuit complexity and neural networks*, The MIT press.

Pearson, K. (1901), "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**(11), 559–572.

Phillips, D. L. (1962), "A technique for the numerical solution of certain integral equations of the first kind," *Journal of the ACM* **9**(1), 84–97.

Pisier, G. (1980–1981), "Remarques sur un résultat non publié de B. maurey."

Pitt, L. & Valiant, L. (1988), "Computational limitations on learning from examples," *Journal of the Association for Computing Machinery* **35**(4), 965–984.

Poon, H. & Domingos, P. (2011), "Sum-product networks: A new deep architecture," in Conference on Uncertainty in Artificial Intelligence (UAI).

Quinlan, J. R. (1986), "Induction of decision trees," *Machine Learning* **1**, 81–106.

Quinlan, J. R. (1993), *C4.5: Programs for machine learning*, Morgan Kaufmann.

Rabiner, L. & Juang, B. (1986), "An introduction to hidden markov models," *IEEE ASSP Magazine* **3**(1), 4–16.

Rakhlin, A., Shamir, O. & Sridharan, K. (2012), "Making gradient descent optimal for strongly convex stochastic optimization," in ICML.

Rakhlin, A., Sridharan, K. & Tewari, A. (2010), "Online learning: Random averages, combinatorial parameters, and learnability," in NIPS.

Rakhlin, S., Mukherjee, S. & Poggio, T. (2005), "Stability results in learning theory," *Analysis and Applications* **3**(4), 397–419.

Ranzato, M., Huang, F., Boureau, Y. & Lecun, Y. (2007), "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE, pp. 1–8.

Rissanen, J. (1978), "Modeling by shortest data description," *Automatica* **14**, 465–471.

Rissanen, J. (1983), "A universal prior for integers and estimation by minimum description length," *The Annals of Statistics* **11**(2), 416–431.

Robbins, H. & Monro, S. (1951), "A stochastic approximation method," *The Annals of Mathematical Statistics*, pp. 400–407.

Downloaded from https://www.cambridge.org/core. Bibliothekssystem Universitaet Hamburg, on 21 Apr 2019 at 16:48:28, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781107298019.036

Rogers, W. & Wagner, T. (1978), "A finite sample distribution-free performance bound for local discrimination rules," *The Annals of Statistics* **6**(3), 506–514.

Rokach, L. (2007), *Data mining with decision trees: Theory and applications*, Vol. 69, World Scientific.

Rosenblatt, F. (1958), "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review* **65**, 386–407. (Reprinted in *Neurocomputing*, MIT Press, 1988).

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986), "Learning internal representations by error propagation," in D. E. Rumelhart & J. L. McClelland, eds, *Parallel distributed processing – explorations in the microstructure of cognition*, MIT Press, chapter 8, pp. 318–362.

Sankaran, J. K. (1993), "A note on resolving infeasibility in linear programs by constraint relaxation," *Operations Research Letters* **13**(1), 19–20.

Sauer, N. (1972), "On the density of families of sets," *Journal of Combinatorial Theory Series A* **13**, 145–147.

Schapire, R. (1990), "The strength of weak learnability," *Machine Learning* **5**(2), 197–227.

Schapire, R. E. & Freund, Y. (2012), *Boosting: Foundations and algorithms*, MIT Press.

Schölkopf, B. & Smola, A. J. (2002), *Learning with kernels: Support vector machines, regularization, optimization and beyond*, MIT Press.

Schölkopf, B., Herbrich, R. & Smola, A. (2001), "A generalized representer theorem," in *Computational learning theory*, pp. 416–426.

Schölkopf, B., Herbrich, R., Smola, A. & Williamson, R. (2000), "A generalized representer theorem," in *NeuroCOLT*.

Schölkopf, B., Smola, A. & Müller, K.-R. (1998), 'Nonlinear component analysis as a kernel eigenvalue problem', *Neural computation* **10**(5), 1299–1319.

Seeger, M. (2003), "Pac-bayesian generalisation error bounds for gaussian process classification," *The Journal of Machine Learning Research* **3**, 233–269.

Shakhnarovich, G., Darrell, T. & Indyk, P. (2006), *Nearest-neighbor methods in learning and vision: Theory and practice*, MIT Press.

Shalev-Shwartz, S. (2007), Online Learning: Theory, Algorithms, and Applications, PhD thesis, The Hebrew University.

Shalev-Shwartz, S. (2011), "Online learning and online convex optimization," *Foundations and Trends® in Machine Learning* **4**(2), 107–194.

Shalev-Shwartz, S., Shamir, O., Srebro, N. & Sridharan, K. (2010), "Learnability, stability and uniform convergence," *The Journal of Machine Learning Research* **9999**, 2635–2670.

Shalev-Shwartz, S., Shamir, O. & Sridharan, K. (2010), "Learning kernel-based halfspaces with the zero-one loss," in COLT.

Shalev-Shwartz, S., Shamir, O., Sridharan, K. & Srebro, N. (2009), "Stochastic convex optimization," in COLT.

Shalev-Shwartz, S. & Singer, Y. (2008), "On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms," in *Proceedings of the nineteenth annual conference on computational learning theory*.

Shalev-Shwartz, S., Singer, Y. & Srebro, N. (2007), "Pegasos: Primal Estimated sub-GrAdient SOlver for SVM," in *International conference on machine learning*, pp. 807–814.

Shalev-Shwartz, S. & Srebro, N. (2008), "SVM optimization: Inverse dependence on training set size," in *International conference on machine learning* ICML, pp. 928–935.

Shalev-Shwartz, S., Zhang, T. & Srebro, N. (2010), "Trading accuracy for sparsity in optimization problems with sparsity constraints," *Siam Journal on Optimization* **20**, 2807–2832.

Downloaded from https://www.cambridge.org/core. Bibliothekssystem Universitaet Hamburg, on 21 Apr 2019 at 16:48:28, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781107298019.036

Shamir, O. & Zhang, T. (2013), "Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes," in ICML.

Shapiro, A., Dentcheva, D. & Ruszczyński, A. (2009), *Lectures on stochastic programming: modeling and theory*, Vol. 9, Society for Industrial and Applied Mathematics.

Shelah, S. (1972), "A combinatorial problem; stability and order for models and theories in infinitary languages," *Pac. J. Math* **4**, 247–261.

Sipser, M. (2006), *Introduction to the Theory of Computation*, Thomson Course Technology.

Slud, E. V. (1977), "Distribution inequalities for the binomial law," *The Annals of Probability* **5**(3), 404–412.

Steinwart, I. & Christmann, A. (2008), *Support vector machines*, Springerverlag, New York.

Stone, C. (1977), "Consistent nonparametric regression," *The Annals of Statistics* **5**(4), 595–620.

Taskar, B., Guestrin, C. & Koller, D. (2003), "Max-margin markov networks," in NIPS.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc B.* **58**(1), 267–288.

Tikhonov, A. N. (1943), "On the stability of inverse problems," *Dolk. Akad. Nauk SSSR* **39**(5), 195–198.

Tishby, N., Pereira, F. & Bialek, W. (1999), "The information bottleneck method," in *The 37'th Allerton conference on communication, control, and computing*.

Tsochantaridis, I., Hofmann, T., Joachims, T. & Altun, Y. (2004), "Support vector machine learning for interdependent and structured output spaces," in *Proceedings of the twenty-first international conference on machine learning*.

Valiant, L. G. (1984), "A theory of the learnable," *Communications of the ACM* **27**(11), 1134–1142.

Vapnik, V. (1992), "Principles of risk minimization for learning theory," in J. E. Moody, S. J. Hanson & R. P. Lippmann, eds., *Advances in Neural Information Processing Systems 4*, Morgan Kaufmann, pp. 831–838.

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, Springer.

Vapnik, V. N. (1982), *Estimation of Dependences Based on Empirical Data*, Springer-Verlag.

Vapnik, V. N. (1998), *Statistical Learning Theory*, Wiley.

Vapnik, V. N. & Chervonenkis, A. Y. (1971), "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and Its Applications* **XVI**(2), 264–280.

Vapnik, V. N. & Chervonenkis, A. Y. (1974), *Theory of pattern recognition*, Nauka, Moscow (In Russian).

Von Luxburg, U. (2007), "A tutorial on spectral clustering," *Statistics and Computing* **17**(4), 395–416.

von Neumann, J. (1928), "Zur theorie der gesellschaftsspiele (on the theory of parlor games)," *Math. Ann.* **100**, 295—320.

Von Neumann, J. (1953), "A certain zero-sum two-person game equivalent to the optimal assignment problem," *Contributions to the Theory of Games* **2**, 5–12.

Vovk, V. G. (1990), "Aggregating strategies," in COLT, pp. 371–383.

Warmuth, M., Glocer, K. & Vishwanathan, S. (2008), "Entropy regularized lpboost," in *Algorithmic Learning Theory* (ALT).

Warmuth, M., Liao, J. & Ratsch, G. (2006), "Totally corrective boosting algorithms that maximize the margin," in *Proceedings of the 23rd international conference on machine learning*.

Downloaded from https://www.cambridge.org/core. Bibliothekssystem Universitaet Hamburg, on 21 Apr 2019 at 16:48:28, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781107298019.036

Weston, J., Chapelle, O., Vapnik, V., Elisseeff, A. & Schölkopf, B. (2002), "Kernel dependency estimation," in Advances in neural information processing systems, pp. 873–880.

Weston, J. & Watkins, C. (1999), "Support vector machines for multi-class pattern recognition," in *Proceedings of the seventh european symposium on artificial neural networks*.

Wolpert, D. H. & Macready, W. G. (1997), "No free lunch theorems for optimization," *Evolutionary Computation, IEEE Transactions on* **1**(1), 67–82.

Zhang, T. (2004), "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *Proceedings of the twenty-first international conference on machine learning*.

Zhao, P. & Yu, B. (2006), "On model selection consistency of Lasso," *Journal of Machine Learning Research* **7**, 2541–2567.

Zinkevich, M. (2003), "Online convex programming and generalized infinitesimal gradient ascent," in *International conference on machine learning*.

Downloaded from https://www.cambridge.org/core. Bibliothekssystem Universitaet Hamburg, on 21 Apr 2019 at 16:48:28, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781107298019.036

Downloaded from https://www.cambridge.org/core. Bibliothekssystem Universitaet Hamburg, on 21 Apr 2019 at 16:48:28, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781107298019.036