[83] H. White. Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, 3:535–549, 1990.

[84] H. White. Learning in artificial neural networks: a statistical perspective. *Neural Computation*, 1(4):425–464, 1990.

[85] K. Yamanishi. A learning criterion for stochastic rules. In *Proceedings of the 3nd Workshop on Computational Learning Theory*, pages 67–81. published by Morgan Kaufmann, 1990.

[69] G. F. Simmons. *Introduction to Topology and Modern Analysis.* McGraw-Hill, New York, 1963.

[70] R. Sloan. Types of noise in data for concept learning. In *Proc. 1988 Workshop on Comp. Learning Theory*, pages 91–96, San Mateo, CA, 1988. Morgan Kaufmann.

[71] H. Sompolinsky, N. Tishby, and H. Seung. Learning from examples in large neural networks. *Phys.Rev.Lett.*, 65:1683–1686, 1990.

[72] N. Tishby, E. Levin, and S. Solla. Consistent infernece of probabilities in layered networks: predictions and generalizations. In *IJCNN International Joint Conference on Neural Networks*, volume II, pages 403–409. IEEE, 1989.

[73] D. Touretsky. *Advances in Neural Information Processing Systems*, volume 1. Morgan Kaufmann, 1989.

[74] D. Touretsky. *Advances in Neural Information Processing Systems*, volume 2. Morgan Kaufmann, 1990.

[75] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, Nov. 1984.

[76] V. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its applications*, XVI(2):264–280, 1971.

[77] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data.* Springer-Verlag, New York, 1982.

[78] V. N. Vapnik. Inductive principles of the search for empirical dependences (methods based on weak convergence of probability measures). In *Proceedings of the 2nd Workshop on Computational Learning Theory*, San Mateo, CA, 1989. published by Morgan Kaufmann.

[79] A. Weigend, B. Huberman, and D. Rumelhart. Predicting the future: A connectionist approach. *International Journal of Neural Systems*, 1:193–209, 1990.

[80] S. Weiss and C. Kulikowski. *Computer Systems that Learn.* Morgan Kaufmann, San Mateo, CA, 1991.

[81] E. Welzl. Partition trees for triangle counting and other range search problems. In *4th ACM Symp. on Comp. Geometry*, pages 23–33, Urbana, IL, 1988.

[82] R. S. Wenocur and R. M. Dudley. Some special Vapnik-Chervonenkis classes. *Discrete Mathematics*, 33:313–318, 1981.

[55] A. Nobel and A. Dembo. On uniform convergence for dependent processes. Technical Report 74, Stanford University, dept. of Statistics, Stanford, CA, 1990.

[56] D. Nolan and D. Pollard. U-processes: Rates of convergence. *Annals of Statistics*, 15(2):780–799, 1987.

[57] S. Nowlan. Maximum likelihood competitive learning. In D. Touretsky, editor, *Advances in Neural Information Processing Systems*, volume 2, pages 574–582. Morgan Kaufmann, 1990.

[58] M. Opper and D. Haussler. Generalization performance of Bayes optimal classification algorithm for learning a perceptron. Technical Report UCSC-CRL-91-3, Computer Research Laboratory, University of Santa Cruz, Feb. 1991.

[59] T. Poggio and F. Girosi. A theory of networks for approximation and learning. Technical Report A.I. Memo No. 1140, Massachusetts Institute of Technology, Cambridge, MA, 1989.

[60] D. Pollard. *Convergence of Stochastic Processes.* Springer-Verlag, 1984.

[61] D. Pollard. Rates of uniform almost-sure convergence for empirical processes indexed by unbounded classes of functions. manuscript, 1986.

[62] D. Pollard. *Empirical Processes: Theory and Applications.* in preparation, 1990.

[63] A. J. Quiroz. Metric entropy and learnability. Unpublished manuscript, Universidad Simón Bolívar, Caracas, Venezuela, 1989.

[64] A. Renyi. *Probability Theory.* North Holland, Amsterdam, 1970.

[65] J. Rissanen. Stochastic complexity and modeling. *The Annals of Statistics*, 14(3):1080–1100, 1986.

[66] D. E. Rumelhart and J. L. McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations.* MIT Press, Cambridge, Mass., 1986.

[67] N. Sauer. On the density of families of sets. *J. Combinatorial Theory (A)*, 13:145–147, 1972.

[68] G. Shackelford and D. Volper. Learning k-DNF with noise in the attributes. In *Proceedings of the 1988 Workshop on Computational Learning Theory*, pages 97–103, San Mateo, CA, 1988. published by Morgan Kaufmann.

[41] S. Kulkarni. On metric entripy, vapnik-chervonenkis dimension, and learnability for a class of distributions. Technical Report LIDS-P-1910, Center for Intelligent Control Systems, MIT, 1989.

[42] S. Kullback. *Information Theory and Statistics.* Wiley, New York, 1959.

[43] Y. LeCun, J. Denker, and S. Solla. Optimal brain damage. In D. Touretsky, editor, *Advances in Neural Information Processing Systems*, volume 2, page 589. Morgan Kaufmann, 1990.

[44] D. V. Lindley. The present position in bayesian statistics. *Statistical Science*, 5(1):44–89, 1990.

[45] N. Lineal, Y. Mansour, and R. Rivest. Results on learnability and the Vapnik-Chervonenkis dimension. In *Proceedings of the 1988 Workshop on Computational Learning Theory*, pages 56–68, San Mateo, CA, 1988. published by Morgan Kaufmann.

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

[47] B. B. Mandelbrot. *The Fractal Geometry of Nature.* W. H. Freeman, 1982.

[48] P. McCullagh and J. A. Nelder. *Generalized Linear Models.* Chapman and Hall, London, 1989.

[49] J. Moody and C. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):281–294, 1989.

[50] K. S. Narendra and M. A. L. Thathachar. *Learning Automata – An Introduction.* Prentice Hall, 1989.

[51] B. K. Natarajan. Learning over classes of distributions. In *Proceedings of the 1988 Workshop on Computational Learning Theory*, pages 408–409, San Mateo, CA, 1988. published by Morgan Kaufmann.

[52] B. K. Natarajan. Probably-approximate learning over classes of distributions. Technical Report HPL-SAL-89-29, Hewlett Packard Labs, Palo Alto, CA, 1989.

[53] B. K. Natarajan. Some results on learning. Technical Report CMU-RI-TR-89-6, Carnegie Mellon, 1989.

[54] B. K. Natarajan and P. Tadepalli. Two new frameworks for learning. In *Proceedings of the 5th International Conference on Machine Learning*, pages 402–415, San Mateo, CA, 1988. published by Morgan Kaufmann.

[27] J. D. Farmer. Information dimension and the probabilistic structure of chaos. *Z. Naturforsch. A*, 37:1304–1325, 1982.

[28] J. D. Farmer, E. Ott, and J. A. Yorke. The dimension of chaotic attractors. *Physica D*, 7:153–180, 1983.

[29] V. Gullapalli. A stochastic reinforcement algortihm for learning real-valued functions. *Neural Networks*, 3(6):671–692, 1990.

[30] G. Gyorgyi and N. Tishby. Statistical theory of learning a rule. In K. Thuemann and R. Koeberle, editors, *Neural Networks and Spin Glasses*. World Scientific, 1990.

[31] D. Haussler. Quantifying inductive bias: AI learning algorithms and Valiant's learning framework. *Artificial Intelligence*, 36:177–221, 1988.

[32] D. Haussler. Learning conjunctive concepts in structural domains. *Machine Learning*, 4:7–40, 1989.

[33] D. Haussler. Decision theoretic generalizations of the pac learning model. In *Proc. First Workshop on Algorithmic Learning Theory*, pages 21–41, Yokyo, Japan, 1990.

[34] D. Haussler, M. Kearns, N. Littlestone, and M. K. Warmuth. Equivalence of models for polynomial learnability. Technical Report UCSC-CRL-88-06, University of Calif. Computer Research Laboratory, Santa Cruz, CA, 1988.

[35] D. Haussler, N. Littlestone, and M. Warmuth. Predicting 0,1-functions on randomly drawn points. Technical Report USCS-CRL-90-54, U.C. Santa Cruz Computer Science Laboratory, 1990.

[36] D. Haussler and E. Welzl. Epsilon nets and simplex range queries. *Disc. Comp. Geometry*, 2:127–151, 1987.

[37] M. Kearns, M. Li, L. Pitt, and L. Valiant. On the learnability of boolean formulae. In *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*, pages 285–295, New York, New York, May 1987.

[38] M. Kearns and R. Schapire. Efficient distribution-free learning of probabilistic concepts. In *31th Annual IEEE Symposium on Foundations of Computer Science*, 1990. to appear.

[39] J. Kiefer. *Introduction to Statistical Inference*. Springer-Verlag, 1987.

[40] A. N. Kolmogorov and V. M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *Amer. Math. Soc. Translations (Ser. 2)*, 17:277–364, 1961.

[13] G. M. Benedek and A. Itai. Learnability by fixed distributions. In *Proc. 1988 Workshop on Comp. Learning Theory*, pages 80–90, San Mateo, CA, 1988. Morgan Kaufmann.

[14] J. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, 1985.

[15] P. Billingsley. *Probability and Measure.* Wiley, New York, 1986.

[16] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *JACM*, 36(4):929–965, 1989.

[17] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth International Group, 1984.

[18] J. Denker, D. Schwartz, B. Wittner, S. Solla, R. Howard, L. Jackel, and J. Hopfield. Automatic learning, rule extraction and generalization. *Complex Syst.*, 1:877–922, 1987.

[19] L. Devroye. Automatic pattern recognition: A study of the probability of error. *IEEE Trans. on Pattern Anal. and Mach. Intelligence*, 10(4):530–543, 1988.

[20] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* Wiley, 1973.

[21] R. M. Dudley. Central limit theorems for empirical measures. *Ann. Prob.*, 6(6):899–929, 1978.

[22] R. M. Dudley. A course on empirical processes. *Lecture Notes in Mathematics*, 1097:2–142, 1984.

[23] R. M. Dudley. Universal Donsker classes and metric entropy. *Ann. Prob.*, 15(4):1306–1326, 1987.

[24] R. Durbin and D. E. Rumelhart. Product units: A computationaly powerful and biologically plausible extension to backporpogation networks. *Neural Computation*, 1(1):133–142, 1989.

[25] H. Edelsbrunner. *Algorithms in Combinatorial Geometry.* Springer-Verlag, 1987.

[26] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. In *Proceedings of the 1988 Workshop on Computational Learning Theory*, pages 139–154, San Mateo, CA, 1988. published by Morgan Kaufmann.

This further simplifies to

$$\frac{16N}{\alpha^2\nu} \geq e \ln \frac{16N}{\alpha^2\nu},$$

which holds, since $x \geq e \ln x$ for any x. Finally, since this inequality holds for the given m, it is easy to see that it will also hold for larger m. $\square$

References

[1] N. Abe and M. Warmuth. On the computational complexity of approximating distributions by probabilistic automata. In *Proceedings of the 3nd Workshop on Computational Learning Theory*, pages 52–66. published by Morgan Kaufmann, 1990.

[2] K. Alexander. Rates of growth for weighted empirical processes. In *Proc. of Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, volume 2, pages 475–493, 1985.

[3] K. Alexander. Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probability Theory and Related Fields*, 75:379–423, 1987.

[4] D. Angluin. Queries and concept learning. *Machine Learning*, 2:319–342, 1988.

[5] D. Angluin and P. D. Laird. Identifying k-cnf formulas from noisy examples. Technical Report YALEU/DCS/TR-478, Yale University Department of Computer Science, June 1986.

[6] D. Angluin and L. G. Valiant. Fast probabilistic algorithms for hamiltonian circuits and matchings. *Journal of Computer and System Sciences*, 18(2):155–193, Apr. 1979.

[7] M. Anthony and J. Shawe-Taylor. A result of Vapnik with applications. Technical Report CSD-TR-628, University of London, Surrey, England, 1990.

[8] P. Assouad. Densité et dimension. *Ann. Inst. Fourier, Grenoble*, 33(3):233–282, 1983.

[9] A. Barron. Statistical properties of artificial neural networks. In *28th Conference on Decision and Control*, pages 280–285, 1989.

[10] A. Barron and T. Cover. Minumum complexity density estimation. *IEEE Transactions on Information Theory*, 1990. to appear.

[11] A. G. Barto and P. Anandan. Pattern recognizing stochastic learning automata. *IEEE Trans. on Systems, Man and Cybernetics*, 15:360–374, 1985.

[12] E. Baum and D. Haussler. What size net gives valid generalization? *Neural Computation*, 1(1):151–160, 1989.

pseudo dimension at most $k+1$. Since the pseudo dimension is the same as the Vapnik Chervonenkis dimension for classes of indicator functions, this implies that the class has Vapnik Chervonenkis dimension at most $k+1$. Now let d be the sum of all the Vapnik Chervonenkis dimensions of all the classes of functions associated with the computation units of the architecture $\mathcal{A}$. It follows that $d \leq W$, the total number of weights and biases in the network.

For each $h \in \mathcal{H}$ let l_h be the loss function associated with h for the discrete loss l, i.e. $l_h(x,y) = 1$ if $y \neq h(x)$, $l_h(x,y) = 0$ if $y = h(x)$. Let $\mathbf{F} = l_{\mathcal{H}} = \{l_h : h \in \mathcal{H}\}$. Let $\bar{\xi} = ((x_1,y_1),\ldots,(x_m,y_m))$ be any fixed sample and $\bar{x} = (x_1,\ldots,x_m)$. It is easily verified that $|\mathbf{F}_{|\bar{\xi}}| = |\mathcal{H}_{|\bar{x}}|$. It is shown in [12], Theorem 1, that for any class $\mathcal{H}$ as above, $|\mathcal{H}_{|\bar{x}}| \leq (Nem/d)^d$ for all $\bar{x} = (x_1,\ldots,x_m)$, where d and N are as above. This implies that

$$|\mathbf{F}_{|\bar{\xi}}| \leq (Nem/d)^d \leq (Nem/W)^W$$

for all samples $\bar{\xi}$ of length m. Since each $l_h \in \mathbf{F}$ is a random variable that is bounded between 0 and 1, Theorem 12 (second part) shows that

$$\mathbf{Pr}\left(\exists h \in \mathbf{F} : d_\nu(\hat{\mathbf{E}}_{\bar{\xi}}(h), \mathbf{E}(h)) > \alpha\right) \leq 4(2eNm/W)^W e^{-\alpha^2\nu m/2}.$$

This gives the first bound.

For the second bound, it can be shown that for sample size

$$m = \frac{4}{\alpha^2\nu}\left(2W\ln\frac{16N}{\alpha^2\nu} + \ln\frac{4}{\delta}\right)$$

we have

$$4(2eNm/W)^W e^{-\alpha^2\nu m/2} \leq \delta.$$

To see this, first note that by rearranging, we get

$$\alpha^2\nu m/2 \geq W\ln(2eNm/W) + \ln(4/\delta).$$

thus it suffices to show

$$\alpha^2\nu m/4 \geq W\ln(2eNm/W) \;\; and \;\; \alpha^2\nu m/4 \geq \ln(4/\delta).$$

The latter inequality is assured by the last term in the formula for m. For the former inequality, let us take m equal to the first term only, i.e.

$$m = \frac{8W}{\alpha^2\nu}\ln\frac{16N}{\alpha^2\nu}.$$

Substituting this into the former inequality and simplifying, we get

$$2\ln\frac{16N}{\alpha^2\nu} \geq \ln\left(\frac{16eN}{\alpha^2\nu}\ln\frac{16N}{\alpha^2\nu}\right).$$

Lemma 14 *For any $m \geq 1$, $\bar{x}, \bar{y} \in (\Re^+)^{2m}$ and $\nu > 0$, $\bar{d}_\nu(\bar{x}, \bar{y}) \leq \frac{2}{\nu} d_{L^1}(\bar{x}, \bar{y})$.*

Proof. For any $\sigma \in \Gamma_{2m}$

$$d_\nu(\mu_1(\bar{x}, \sigma), \mu_1(\bar{y}, \sigma)) + d_\nu(\mu_2(\bar{x}, \sigma), \mu_2(\bar{y}, \sigma))$$

$$= \frac{|\sum_{i=1}^{m}(x_{\sigma(i)} - y_{\sigma(i)})|}{\nu m + \sum_{i=1}^{m}(x_{\sigma(i)} + y_{\sigma(i)})} + \frac{|\sum_{i=m+1}^{2m}(x_{\sigma(i)} - y_{\sigma(i)})|}{\nu m + \sum_{i=m+1}^{2m}(x_{\sigma(i)} + y_{\sigma(i)})}$$

$$\leq \frac{\sum_{i=1}^{2m}|x_{\sigma(i)} - y_{\sigma(i)}|}{\nu m} = \frac{2}{\nu} d_{L^1}(\bar{x}, \bar{y}).$$

The result follows. □

The theorem follows easily from the last two lemmas. □

10.5 Bounds on sample size for learning in feedforward nets with sharp thresholds

In this section we give the bounds for uniform convergence of empirical estimates in neural networks with sharp threshold functions claimed in section 7.

Theorem 13 *Let $\mathcal{A}$ be a feedforward architecture as defined in section 7 with $n \geq 1$ inputs, one output, $N \geq 2$ computation units and a total of W weights and biases. Let $X = \Re^n$, $Y = A = \{0, 1\}$, and l be the discrete loss function. Assume that the squashing function for each computation unit has the form $sign \circ \sigma$, where σ is non-decreasing or non-increasing. Different σ's can be used for different units. Let $\mathcal{H}$ be all functions from X into A representable on $\mathcal{A}$ by varying the weights and biases. Let $\nu > 0$, $0 < \alpha < 1$ and $m \geq 1$. Suppose that $\bar{\xi}$ is generated by m independent random draws from a probability measure P on $S = X \times Y$. Then*

$$\mathbf{Pr}\left(\exists h \in \mathcal{H} : d_\nu(\widehat{\mathbf{er}}_{\bar{\xi},l}(h), \mathbf{er}_{P,l}(h)) > \alpha\right) \leq 4(2eNm/W)^W e^{-\alpha^2 \nu m/2}.$$

This probability is at most δ for a sample size m that is

$$O\left(\frac{1}{\alpha^2 \nu}\left(W \log \frac{N}{\alpha\nu} + \log \frac{1}{\delta}\right)\right).$$

Proof. Each computation unit in the network with k weights is associated with a class of $\{0, 1\}$-valued functions of the form

$$f(\bar{x}) = sign(\sigma(\phi_1(\bar{x}), \ldots, \phi_k(\bar{x})) + \theta + \sum_{j=1}^{k} w_j \phi_j(\bar{x})))$$

where $\theta, w_1, \ldots, w_k$ are adjustable real-valued parameters and ϕ_1, ..., ϕ_k, μ and σ are fixed functions, the latter monotone. By Theorems 3 and 4 this class of functions has

Thus if we draw $\bar{\xi}$ at random from S^{2m} and independently select a permutation $\sigma \in \Gamma_{2m}$ uniformly at random,

$$\mathbf{Pr}\left(\exists f \in \mathbf{F}\ with\ d_\nu(\hat{\mathbf{E}}'_{\sigma(\bar{\xi})}(f), \hat{\mathbf{E}}''_{\sigma(\bar{\xi})}(f)) > \alpha/2\right) \leq \mathbf{E}(min(2|\mathbf{F}_{|\bar{\xi}}|e^{-\alpha^2\nu m/2M}, 1)).$$

However, since each of the $2m$ observations in $\bar{\xi}$ are independent, each of the samples $\sigma(\bar{\xi})$ for $\sigma \in \Gamma_{2m}$ are equally likely. Hence

$$\mathbf{Pr}\left(\exists f \in \mathbf{F}\ with\ d_\nu(\hat{\mathbf{E}}'_{\sigma(\bar{\xi})}(f), \hat{\mathbf{E}}''_{\sigma(\bar{\xi})}(f)) > \alpha/2\right),$$

where both $\bar{\xi}$ and σ are chosen at random, is the same as

$$\mathbf{Pr}\left(\exists f \in \mathbf{F}\ with\ d_\nu(\hat{\mathbf{E}}'_{\bar{\xi}}(f), \hat{\mathbf{E}}''_{\bar{\xi}}(f)) > \alpha/2\right),$$

where only the sample $\bar{\xi}$ is chosen at random. The second bound follows.

The proof of the first bound is similar, except for steps 7 and 8. From Lemma 10, using the extension $\bar{d}_\nu$ of the metric d_ν to even-length sequences of reals given in Definition 12 above, if $f, f^* \in \mathbf{F}$ are such that

$$\bar{d}_\nu((f(\xi_1), \ldots, f(\xi_{2m})), (f^*(\xi_1), \ldots, f^*(\xi_{2m}))) \leq \alpha/4,$$

then for any $\sigma \in \Gamma_{2m}$ such that

$$d_\nu(\hat{\mathbf{E}}'_{\sigma(\bar{\xi})}(f), \hat{\mathbf{E}}''_{\sigma(\bar{\xi})}(f)) > \alpha/2,$$

we must have

$$d_\nu(\hat{\mathbf{E}}'_{\sigma(\bar{\xi})}(f^*), \hat{\mathbf{E}}''_{\sigma(\bar{\xi})}(f^*)) > \alpha/4.$$

Thus if N is an $\alpha/4$-cover for $\mathbf{F}_{|\bar{\xi}}$ with respect to the metric $\bar{d}_\nu$, then whenever there exists $f \in \mathbf{F}$ with

$$d_\nu(\hat{\mathbf{E}}'_{\sigma(\bar{\xi})}(f), \hat{\mathbf{E}}''_{\sigma(\bar{\xi})}(f)) > \alpha/2,$$

there exists $f^* \in N$ with

$$d_\nu(\hat{\mathbf{E}}'_{\sigma(\bar{\xi})}(f^*), \hat{\mathbf{E}}''_{\sigma(\bar{\xi})}(f^*)) > \alpha/4.$$

For fixed f^* and random σ, the probability of the latter event is at most $2e^{-\alpha^2\nu m/8M}$ by Lemma 11. Hence for any fixed $\bar{\xi} \in S^{2m}$, if we select a permutation $\sigma \in \Gamma_{2m}$ uniformly at random,

$$\mathbf{Pr}\left(\exists f \in \mathbf{F}\ with\ d_\nu(\hat{\mathbf{E}}'_{\sigma(\bar{\xi})}(f), \hat{\mathbf{E}}''_{\sigma(\bar{\xi})}(f)) > \alpha/2\right) \leq min(2\mathcal{N}(\alpha/4, \mathbf{F}_{|\bar{\xi}}, \bar{d}_\nu)e^{-\alpha^2\nu m/8M}, 1)).$$

The remainder of the proof is as above. □

We need only one more lemma to complete our proof; one that can be used to relate the $\bar{d}_\nu$ covering numbers to the d_{L^1} covering numbers.

sample $\bar{\xi}'' \in S^m$, then with probability at least $1/2$, $d_\nu(\hat{\mathbf{E}}_{\bar{\xi}''}(f), \mathbf{E}(f)) \leq \alpha/2$. Whenever this happens we have $d_\nu(\hat{\mathbf{E}}_{\bar{\xi}'}(f), \hat{\mathbf{E}}_{\bar{\xi}''}(f)) > \alpha/2$ by the triangle inequality for d_ν. Thus

$$\begin{aligned}
& \mathbf{Pr}\left\{\bar{\xi} \in S^{2m} : \exists f \in \mathbf{F} \text{ with } d_\nu(\hat{\mathbf{E}}'_{\bar{\xi}}(f), \hat{\mathbf{E}}''_{\bar{\xi}}(f)) > \alpha/2\right\} \\
\geq\ & \mathbf{Pr}\left\{\bar{\xi}'\bar{\xi}'' \in S^{2m} : \exists f \in \mathbf{F} \text{ with } d_\nu(\hat{\mathbf{E}}_{\bar{\xi}'}(f), \mathbf{E}(f)) > \alpha \text{ and } d_\nu(\hat{\mathbf{E}}_{\bar{\xi}''}(f), \mathbf{E}(f)) \leq \alpha/2\right\} \\
\geq\ & \frac{1}{2}\mathbf{Pr}\left\{\bar{\xi}' \in S^m : \exists f \in \mathbf{F} \text{ with } d_\nu(\hat{\mathbf{E}}_{\bar{\xi}'}(f), \mathbf{E}(f)) > \alpha\right\}.
\end{aligned}$$

Again, when S and $\mathbf{F}$ are uncountable, permissibility guarantees that the implied use of Fubini's theorem in obtaining the above inequalities is justified. □

We are now in a position to prove the following version of our theorem, using the extended metric $\bar{d}_\nu$ in place of the L^1 metric to measure covering numbers.

Lemma 13 *Let $\mathbf{F}$ be a permissible set of functions on S with $0 \leq f(\xi) \leq M$ for all $f \in \mathbf{F}$ and $\xi \in S$. Assume $\nu > 0$, $0 < \alpha < 1$ and $m \geq 1$. Let*

$$p(\alpha, \nu, m) = \mathbf{Pr}\left\{\bar{\xi} \in S^m : \exists f \in \mathbf{F} \text{ with } d_\nu(\hat{\mathbf{E}}_{\bar{\xi}}(f), \mathbf{E}(f)) > \alpha\right\}.$$

Then

$$p(\alpha, \nu, m) \leq 2\mathbf{E}(min(2\mathcal{N}(\alpha/4, \mathbf{F}_{|\bar{\xi}}, \bar{d}_\nu)e^{-\alpha^2\nu m/8M}, 1)),$$

and if in addition $\mathbf{F}_{|\bar{\xi}}$ is finite for all $\bar{\xi} \in S^{2m}$ then

$$p(\alpha, \nu, m) \leq 2\mathbf{E}(min(2|\mathbf{F}_{|\bar{\xi}}|e^{-\alpha^2\nu m/2M}, 1)),$$

where the expectations are over $\bar{\xi}$ drawn randomly from S^{2m}.

Proof. First note that both bounds are trivial if $m < 2M/(\alpha^2\nu)$, so we may assume $m \geq 2M/(\alpha^2\nu)$. Hence by Lemma 12,

$$p(\alpha, \nu, m) \leq 2\mathbf{Pr}\left\{\bar{\xi} \in S^{2m} : \exists f \in \mathbf{F} \text{ with } d_\nu(\hat{\mathbf{E}}'_{\bar{\xi}}(f), \hat{\mathbf{E}}''_{\bar{\xi}}(f)) > \alpha/2\right\}.$$

Thus it suffices to obtain bounds for the latter quantity.

We begin with the second bound, for the case when $\mathbf{F}_{|\bar{\xi}}$ is always finite. For any sample $\bar{\xi} = (\xi_1, \ldots, \xi_{2m}) \in S^{2m}$ and $\sigma \in \Gamma_{2m}$, let $\sigma(\bar{\xi}) = (\xi_{\sigma(1)}, \ldots, \xi_{\sigma(2m)})$. For any fixed function $f \in \mathbf{F}$ and fixed $\bar{\xi} \in S^{2m}$, if we select a permutation $\sigma \in \Gamma_{2m}$ uniformly at random,

$$\mathbf{Pr}\left(d_\nu(\hat{\mathbf{E}}'_{\sigma(\bar{\xi})}(f), \hat{\mathbf{E}}''_{\sigma(\bar{\xi})}(f)) > \alpha/2\right) \leq 2e^{-\alpha^2\nu m/2M} \tag{7}$$

by Lemma 11. Hence for any fixed $\bar{\xi} \in S^{2m}$, if we select a permutation $\sigma \in \Gamma_{2m}$ uniformly at random,

$$\mathbf{Pr}\left(\exists f \in \mathbf{F} \text{ with } d_\nu(\hat{\mathbf{E}}'_{\sigma(\bar{\xi})}(f), \hat{\mathbf{E}}''_{\sigma(\bar{\xi})}(f)) > \alpha/2\right) \leq min(2|\mathbf{F}_{|\bar{\xi}}|e^{-\alpha^2\nu m/2M}, 1). \tag{8}$$

because each swap in a randomly chosen $\sigma \in \Gamma_{2m}$ is independent. Since $\mathbf{E}(Y_i) = 0$ and $-|x_i - x_{m+i}| \leq Y_i \leq |x_i - x_{m+i}|$, we can apply Hoeffding's inequality (see e.g. [60]) to bound the latter probability by

$$2e^{-\alpha^2(\nu m + \sum_{i=1}^{2m} x_i)^2/2\sum_{i=1}^{m}(x_i - x_{m+i})^2}.$$

Let $\beta = \sum_{i=1}^{2m} x_i$. Since $0 \leq x_i \leq M$,

$$\sum_{i=1}^{m}(x_i - x_{m+i})^2 \leq \sum_{i=1}^{m} M|x_i - x_{m+i}| \leq \beta M.$$

Hence we have

$$2e^{-\alpha^2(\nu m + \sum_{i=1}^{2m} x_i)^2/2\sum_{i=1}^{m}(x_i - x_{m+i})^2} \leq 2e^{-\alpha^2(\nu m+\beta)^2/2\beta M} = 2e^{-\frac{\alpha^2}{2M}(\frac{(\nu m+\beta)^2}{\beta})}.$$

The expression in parentheses is minimized, and therefore the whole expression is maximized, by setting $\beta = \nu m$, giving a value of $4\nu m$. Hence

$$\mathbf{Pr}\left(d_\nu(\mu_1(\bar{x}, \sigma), \mu_2(\bar{x}, \sigma)) > \alpha\right) \leq 2e^{-2\alpha^2\nu m/M}.$$

□

For our next lemma we will need some notation to refer to the separate empirical estimates based on the first and second halves of an even length sample.

Definition 13 *For all $m \geq 1$ and $\bar{\xi} \in S^{2m}$, we let $\widehat{\mathbf{E}}'_{\bar{\xi}}(f) = \frac{1}{m}\sum_{i=1}^{m} f(\xi_i)$ and $\widehat{\mathbf{E}}''_{\bar{\xi}}(f) = \frac{1}{m}\sum_{i=m+1}^{2m} f(\xi_i)$.*

Lemma 12 *Let $\mathbf{F}$ be a permissible set of functions on S with $0 \leq f(\xi) \leq M$ for all $f \in \mathbf{F}$ and $\xi \in S$. Assume $\nu > 0$, $0 < \alpha < 1$ and $m \geq 2M/(\alpha^2\nu)$. Then*

$$\begin{aligned} &\mathbf{Pr}\left\{\bar{\xi} \in S^m : \exists f \in \mathbf{F} \text{ with } d_\nu(\widehat{\mathbf{E}}_{\bar{\xi}}(f), \mathbf{E}(f)) > \alpha\right\} \\ \leq\ &2\mathbf{Pr}\left\{\bar{\xi} \in S^{2m} : \exists f \in \mathbf{F} \text{ with } d_\nu(\widehat{\mathbf{E}}'_{\bar{\xi}}(f), \widehat{\mathbf{E}}''_{\bar{\xi}}(f)) > \alpha/2\right\}. \end{aligned}$$

Proof. If S and $\mathbf{F}$ are uncountable, the assumption of permissibility guarantees that these probabilities are well-defined (see appendix; [60]). From Chebyshev's inequality (see Lemma 9, part (1) in the Appendix), for each individual $f \in \mathbf{F}$,

$$\mathbf{Pr}\left\{\bar{\xi} \in S^m : d_\nu(\widehat{\mathbf{E}}_{\bar{\xi}}(f), \mathbf{E}(f)) > \alpha/2\right\} \leq M/(\alpha^2\nu m).$$

Since $m \geq 2M/(\alpha^2\nu)$, this probability is at most $1/2$. Now consider any $f \in \mathbf{F}$ and sample $\bar{\xi}' \in S^m$ such that $d_\nu(\widehat{\mathbf{E}}_{\bar{\xi}'}(f), \mathbf{E}(f)) > \alpha$. If we draw an independent random

It is easily verified that $\bar{d}$ is a pseudo-metric on $(\Re^+)^{2m}$. Symmetry is obvious, and the triangle inequality follows easily from the triangle inequality for d on $\Re^+$:

$$\begin{aligned}
\bar{d}(\bar{x},\bar{y}) + \bar{d}(\bar{y},\bar{z}) &= max\{d(\mu_1(\bar{x},\sigma),\mu_1(\bar{y},\sigma)) + d(\mu_2(\bar{x},\sigma),\mu_2(\bar{y},\sigma)) : \sigma \in \Gamma_{2m}\} \\
&\quad + max\{d(\mu_1(\bar{y},\sigma),\mu_1(\bar{z},\sigma)) + d(\mu_2(\bar{y},\sigma),\mu_2(\bar{z},\sigma)) : \sigma \in \Gamma_{2m}\} \\
&\geq max\{d(\mu_1(\bar{x},\sigma),\mu_1(\bar{y},\sigma)) + d(\mu_2(\bar{x},\sigma),\mu_2(\bar{y},\sigma)) \\
&\quad + d(\mu_1(\bar{y},\sigma),\mu_1(\bar{z},\sigma)) + d(\mu_2(\bar{y},\sigma),\mu_2(\bar{z},\sigma)) : \sigma \in \Gamma_{2m}\} \\
&\geq max\{d(\mu_1(\bar{x},\sigma),\mu_1(\bar{z},\sigma)) + d(\mu_2(\bar{x},\sigma),\mu_2(\bar{z},\sigma)) : \sigma \in \Gamma_{2m}\} \\
&= \bar{d}(\bar{x},\bar{z})
\end{aligned}$$

We note the following additional property of this extension.

Lemma 10 *For all $\bar{x},\bar{y} \in (\Re^+)^{2m}$ and $\sigma \in \Gamma_{2m}$, $d(\mu_1(\bar{x},\sigma),\mu_2(\bar{x},\sigma)) \leq d(\mu_1(\bar{y},\sigma),\mu_2(\bar{y},\sigma)) + \bar{d}(\bar{x},\bar{y})$.*

Proof. We have

$$d(\mu_1(\bar{x},\sigma),\mu_2(\bar{x},\sigma)) \leq d(\mu_1(\bar{y},\sigma),\mu_2(\bar{y},\sigma)) + d(\mu_1(\bar{x},\sigma),\mu_1(\bar{y},\sigma)) + d(\mu_2(\bar{x},\sigma),\mu_2(\bar{y},\sigma))$$

by the triangle inequality on d. The last two terms of this sum combined are at most $\bar{d}(\bar{x},\bar{y})$ by definition. □

We now restrict ourselves to the case that the metric d is the metric d_ν for some $\nu > 0$. The following lemma will play a key role in establishing our basic exponential inequality.

Lemma 11 *Let $\bar{x} = (x_1,\ldots,x_{2m})$ be a sequence of reals such that $0 \leq x_i \leq M$, $1 \leq i \leq 2m$. Assume $\nu > 0$ and $0 < \alpha < 1$. Then if a permutation $\sigma \in \Gamma_{2m}$ is chosen uniformly at random,*

$$\mathbf{Pr}\left(d_\nu(\mu_1(\bar{x},\sigma),\mu_2(\bar{x},\sigma)) > \alpha\right) \leq 2e^{-2\alpha^2\nu m/M}$$

Proof. For each i, $1 \leq i \leq m$, let Y_i be an independent random variable such that $Y_i = x_i - x_{m+i}$ with probability 1/2 and $Y_i = x_{m+i} - x_i$ with probability 1/2. Note that for any $\sigma \in \Gamma_{2m}$,

$$d_\nu(\mu_1(\bar{x},\sigma),\mu_2(\bar{x},\sigma)) = \frac{|\frac{1}{m}\sum_{i=1}^{m} x_{\sigma(i)} - \frac{1}{m}\sum_{i=m+1}^{2m} x_{\sigma(i)}|}{\nu + \frac{1}{m}\sum_{i=1}^{2m} x_{\sigma(i)}} = \frac{|\sum_{i=1}^{m}(x_{\sigma(i)} - x_{\sigma(m+i)})|}{\nu m + \sum_{i=1}^{2m} x_i}.$$

Hence

$$\begin{aligned}
&\mathbf{Pr}\left(d_\nu(\mu_1(\bar{x},\sigma),\mu_2(\bar{x},\sigma)) > \alpha\right) \\
&= \mathbf{Pr}\left(|\sum_{i=1}^{m}(x_{\sigma(i)} - x_{\sigma(m+i)})| > \alpha(\nu m + \sum_{i=1}^{2m} x_i)\right) \\
&= \mathbf{Pr}\left(|\sum_{i=1}^{m} Y_i| > \alpha(\nu m + \sum_{i=1}^{2m} x_i)\right),
\end{aligned}$$

10.4 Proof of Main Theorem on Uniform Convergence of Empirical Estimates

In this section we prove Theorem 2 from section 3. Actually, we will state and prove a slightly stronger result. This result is obtained by bounding the probability of uniform convergence on a sample of length m in terms of the expected covering numbers associated with a sample of length $2m$, and by expanding the expectation to include the negative exponential term with a "truncation" at 1. It turns out that this saves us a factor of $1/2$ in the negative exponential term. We also include special bounds for the case that $\mathbf{F}_{|\bar{\xi}}$ is always finite. This case comes up, for example, when $\mathbf{F} = l_{\mathcal{H}}$ and we use the discrete loss function l, as in the PAC learning model.

Theorem 12 *Let $\mathbf{F}$ be a permissible set of functions on S with $0 \leq f(\xi) \leq M$ for all $f \in \mathbf{F}$ and $\xi \in S$. Assume $\nu > 0$, $0 < \alpha < 1$ and $m \geq 1$. Suppose that $\bar{\xi}$ is generated by m independent random draws according to any probability measure on S. Let*

$$p(\alpha, \nu, m) = \mathbf{Pr}\left\{\bar{\xi} \in S^m : \exists f \in \mathbf{F} \text{ with } d_\nu(\hat{\mathbf{E}}_{\bar{\xi}}(f), \mathbf{E}(f)) > \alpha\right\}.$$

Then

$$p(\alpha, \nu, m) \leq 2\mathbf{E}(\min(2\mathcal{N}(\alpha\nu/8, \mathbf{F}_{|\bar{\xi}}, d_{L^1})e^{-\alpha^2\nu m/8M}, 1)),$$

where the expectation is over $\bar{\xi}$ drawn randomly from S^{2m}. If in addition $\mathbf{F}_{|\bar{\xi}}$ is finite for all $\bar{\xi} \in S^{2m}$ then

$$p(\alpha, \nu, m) \leq 2\mathbf{E}(min(2|\mathbf{F}_{|\bar{\xi}}|e^{-\alpha^2\nu m/2M}, 1)).$$

Theorem 2 is obtained as a corollary of this result by substituting $m/2$ for m and not taking the minimum with 1 in the left hand side of the first bound for $p(\alpha, \nu, m)$.

The proof is given in a series of lemmas. We first extend the metric d_ν to a pseudo-metric on vectors in $(\Re^+)^{2m}$ for $m \geq 1$. We will do so in a somewhat unusual manner that will be useful in what follows. In fact, this extension can be defined for any metric, so we will state it in its general form.

Definition 12 *For each integer $m \geq 1$, let Γ_{2m} denote the set of all permutations σ of $\{1, \ldots, 2m\}$ such that for all i, $1 \leq i \leq m$, either $\sigma(i) = m + i$ and $\sigma(m + i) = i$, or $\sigma(i) = i$ and $\sigma(m + i) = m + i$. Thus the permutations in Γ_{2m} swap selected indices in the first half of the sequence $\{1, \ldots, 2m\}$ with corresponding indices in the second half. For any $\bar{x} = (x_1, \ldots, x_{2m}) \in (\Re^+)^{2m}$ and $\sigma \in \Gamma_{2m}$, let $\mu_1(\bar{x}, \sigma) = \frac{1}{m}\sum_{i=1}^{m} x_{\sigma(i)}$ and $\mu_2(\bar{x}, \sigma) = \frac{1}{m}\sum_{i=m+1}^{2m} x_{\sigma(i)}$. For any metric d on $\Re^+$ and $\bar{x}, \bar{y} \in (\Re^+)^{2m}$, let $\bar{d}(\bar{x}, \bar{y}) = max\{d(\mu_1(\bar{x}, \sigma), \mu_1(\bar{y}, \sigma)) + d(\mu_2(\bar{x}, \sigma), \mu_2(\bar{y}, \sigma)) : \sigma \in \Gamma_{2m}\}$.*

2.

$$\mathbf{Pr}\left(d_\nu(\frac{1}{n}\sum_{i=1}^{n} Z_i, \mu) > \alpha\right) \leq 2e^{\frac{-18\alpha^2\nu n}{(3+\alpha)^2 M}} < 2e^{-\frac{9}{8}\alpha^2\nu n/M} < 2e^{-\alpha^2\nu n/M}.$$

Proof. Let $Y_i = Z_i - \mu$, $1 \leq i \leq n$. Then

$$\begin{aligned} & \mathbf{Pr}\left(d_\nu(\frac{1}{n}\sum_{i=1}^{n} Z_i, \mu) > \alpha\right) \\ = \; & \mathbf{Pr}\left(\frac{|(\sum_{i=1}^{n} Z_i) - \mu n|}{\nu n + \sum_{i=1}^{n} Z_i + \mu n} > \alpha\right) \\ \leq \; & \mathbf{Pr}\left(|\sum_{i=1}^{n} Y_i| > \alpha n(\nu+\mu)\right), \end{aligned}$$

since $\sum_{i=1}^{n} Z_i \geq 0$.

To obtain the first bound, note that by Chebyshev's inequality,

$$\mathbf{Pr}\left(|\sum_{i=1}^{n} Y_i| > \alpha n(\nu+\mu)\right) \leq \sigma^2/(\alpha n(\nu+\mu))^2,$$

where $\sigma^2 = n\mathbf{Var}(Y_i)$. Since $0 \leq Z_i \leq M$, $\mathbf{Var}(Z_i) = \mathbf{Var}(Y_i) \leq \mu(M-\mu)$. Hence

$$\mathbf{Pr}\left(|\sum_{i=1}^{n} Y_i| > \alpha n(\nu+\mu)\right) \leq \frac{\mu(M-\mu)}{\alpha^2 n(\nu+\mu)^2}.$$

It is easily verified that the maximum value of this expression occurs at $\mu = (\nu M)/(2\nu+M)$, and that this gives an upper bound of

$$\frac{M^2}{4\alpha^2\nu n(M+\nu)}.$$

To obtain the second bound, we apply Bernstein's inequality (see e.g. [60], Page 192), which states that

$$\mathbf{Pr}\left(|\sum_{i=1}^{n} Y_i| > \eta\right) \leq 2e^{-\eta^2/2(n\mathbf{Var}(Y_i)+\frac{1}{3}B\eta)}$$

for any zero mean i.i.d. random variables $Y_1, \ldots, Y_n$ bounded in absolute value by B. Substituting $\eta = \alpha n(\nu+\mu)$ and upper bounds $B \leq M$ and $\mathbf{Var}(Y_i) \leq \mu M$, this gives a bound of

$$2e^{-\alpha^2 n^2(\nu+\mu)^2/2(n\mu M + \frac{1}{3}M\alpha n(\nu+\mu))} = 2e^{\frac{-3\alpha^2 n(\nu+\mu)^2}{2M(\alpha\nu+(3+\alpha)\mu)}}.$$

Since $(\nu+\mu)^2/(\alpha\nu+(3+\alpha)\mu)$ is minimized at $\mu = \frac{3-\alpha}{3+\alpha}\nu$, the latter expression is bounded by substituting this value of μ. This gives the first bound of part (2). The second bound follows from the fact that $\alpha < 1$. □

on $\bar{T}$, i.e. if T is in the smallest σ-algebra on $\bar{T}$ containing the open sets. The σ-algebra $\mathcal{B}(T)$ of Borel sets on T is then the restriction to T of the σ-algebra $\mathcal{B}$ of Borel sets on $\bar{T}$, i.e.

$$\mathcal{B}(T) = \{B \cap T : B \in \mathcal{B}\}.$$

Finally, if A is any σ-algebra on S and $\mathcal{C}$ is any σ-algebra on T then $\mathcal{A} \times \mathcal{C}$ denotes the smallest σ algebra on $S \times T$ that contains $\{A \times C : A \in \mathcal{A}, C \in \mathcal{C}\}$. We are now ready for our main definition.

Definition 11 *We say that the class* $\mathbf{F}$ *is* permissible *if it can be indexed by a set* T *such that*

1. T *is a Borel subspace of a compact metric space* $\bar{T}$ *and*
2. *the function* $f : S \times T \to \Re$ *that indexes* $\mathbf{F}$ *by* T *is measurable with respect to the* σ*-algebra* $\mathcal{A} \times \mathcal{B}(T)$.

Most uncountable classes of functions that come up in practice can be indexed by a finite number of real parameters (i.e. with $T = \Re^n$ for some $n \geq 1$) in such a way that condition (2) is satisfied. Condition (1) is satisfied as well in this case, since we can take $\bar{T}$ to be the one-point compactification of T, obtained by adding a *point at infinity* to T (see e.g. [69]).

Results given in Pollard ([60], Appendix C) imply that the sets used in Lemmas 12 and 13 are measurable when $\mathbf{F}$ is permissible. He also shows that the packing numbers $\mathcal{M}(\epsilon, \mathbf{F}_{|\bar{\xi}}, d_{L^1})$ are measurable functions of $\bar{\xi} \in S^m$ for any $m \geq 1$. Since Theorem 11 relates these closely to the covering numbers $\mathcal{N}(\epsilon, \mathbf{F}_{|\bar{\xi}}, d_{L^1})$, this allows us to further formalize our usage of random covering numbers. A more formal treatment would either replace the covering numbers with the packing numbers in our upper bounds, or reword probabilistic bounds on the covering numbers to use outer measure arguments.

10.3 Measuring the accuracy of empirical estimates with the d_ν metric

In this section we give two bounds on the probability of large deviation of empirical estimates from true means, as measured by the d_ν metric. One is derived from Chebyshev's inequality and the other from Bernstein's inequality. The first bound is better for estimates obtained from small samples, the latter for estimates obtained from larger samples.

Lemma 9 *Let* $Z_1, \ldots, Z_n$ *be i.i.d. random variables with range* $0 \leq Z_i \leq M$ *and* $\mathbf{E}(Z_i) = \mu$, $1 \leq i \leq n$. *Assume* $\nu > 0$ *and* $0 < \alpha < 1$. *Then*

1.

$$\mathbf{Pr}\left(d_\nu(\frac{1}{n}\sum_{i=1}^{n} Z_i, \mu) > \alpha\right) \leq \frac{M^2}{4\alpha^2\nu n(M+\nu)} < \frac{M}{4\alpha^2\nu n}.$$

refer to $\mathcal{M}(\epsilon, T, \rho)$ as a *packing number*. The third argument to $\mathcal{N}$ and $\mathcal{M}$ will be omitted when the metric ρ is clear from the context.

The following inequalities are easily verified (see e.g. [40]):

Theorem 11 *If T is a totally bounded subset of the (pseudo) metric space (S, ρ) then for any $\epsilon > 0$,*

$$\mathcal{M}(2\epsilon, T, \rho) \leq \mathcal{N}(\epsilon, T, \rho) \leq \mathcal{M}(\epsilon, T, \rho).$$

Hence both these measures of boundedness, by covering number and by packing number, are equivalent to within a factor of 2 of ϵ. Following [40] we define the *upper metric dimension* of a (pseudo) metric space (S, ρ) by

$$\overline{\mathbf{dim}}(S) = \mathbf{limsup}_{\epsilon \to 0} \frac{\log \mathcal{N}(\epsilon, S, \rho)}{\log(1/\epsilon)}.$$

The *lower metric dimension*, denoted by $\underline{\mathbf{dim}}$, of a (pseudo) metric space (S, ρ) is defined similarly using **liminf**. When $\overline{\mathbf{dim}}(S) = \underline{\mathbf{dim}}(S)$, then this quantity is denoted $\mathbf{dim}(S)$, and referred to simply as the *metric dimension* of (S, ρ). This quantity has also been called the *fractal dimension* [27] and the *capacity dimension* [28]. A very lucid and intuitive treatment is given in [47].

10.2 Permissible classes of functions

In order to obtain the uniform convergence results given in Theorem 2, certain measurability assumptions have to be made concerning the class of functions $\mathbf{F}$ when this class is uncountable. These we have indicated by saying that $\mathbf{F}$ must be a *permissible* class [60]. Here we give a definition of permissible that is a special case of that given by Pollard. This definition will be suitable for our purposes; we refer the reader to [60] and [22] for a more general treatment. See exercise 10, page 39 of [60] for an indication of the kind of problems that can come up with non-permissible classes.

Throughout the paper we have assumed that $\mathbf{F}$ is a class of real-valued functions on a set S, and that P is a measure defined on some σ-algebra $\mathcal{A}$ of subsets of S such that each function in $\mathbf{F}$ is measurable. We will need further conditions on $\mathbf{F}$ when it is uncountable. Let us say that the class $\mathbf{F}$ is *indexed* by the set T if

$$\mathbf{F} = \{f(\cdot, t) : t \in T\},$$

where f is a real-valued function on $S \times T$ and $f(\cdot, t)$ denotes the real-valued function on S obtained from f by fixing the second parameter to t. We will say that the function f *indexes* $\mathbf{F}$ *by* T.

We will need some structure on T as well. If T is contained in a topological space $\bar{T}$ then we will say that T is a *Borel subspace* if T is a Borel set with respect to the topology

bounds, practitioners may still need to rely at least in part on empirically derived sample size bounds.

Finally, many other issues would need to be considered in a complete treatment of the problem of overfitting, including distribution specific bounds on sample complexity (Theorem 2 is actually distribution specific, since the random covering numbers are distribution specific, yet we only apply it here in a distribution independent setting), hypothesis spaces with infinite pseudo and metric dimensions (these include various classes of "smooth" functions and their relatives, see [22], Chapter 7 and [63]) and non i.i.d. sources of examples (see e.g. [83,55]). Despite these shortcomings, we feel that the theory we give here provides useful insights into the nature of the problem of overfitting in learning, and because of its generality will be a useful starting point for further research in this area.

9 Acknowledgements

I would like to thank Dana Angluin, David Pollard and Phil Long for their careful criticisms of an earlier draft of this paper, and their numerous suggestions for improvements. I also thank Naoki Abe, Anselm Blumer, Richard Dudley, and Michael Kearns for helpful comments on earlier drafts. I would also like to thank Ron Rivest, David Rumelhart, Andrzej Ehrenfeucht and Nick Littlestone for stimulating discussions on these topics.

10 Appendix

10.1 Metric spaces, covering numbers and metric dimension

A *pseudo metric* on a set S is a function ρ from $S \times S$ into $\Re^+$ such that for all $x, y, z \in S$, $x = y \Rightarrow \rho(x,y) = 0$, $\rho(x,y) = \rho(y,x)$ (symmetry), and $\rho(x,z) \leq \rho(x,y) + \rho(y,z)$ (triangle inequality). If in addition $\rho(x,y) = 0 \Rightarrow x = y$, then ρ is a *metric*. (S,ρ) is a *(pseudo) metric space*. (S,ρ) is *complete* if every Cauchy sequence of points in S converges to a point in S; (S,ρ) is *separable* if it contains a countable dense subset, i.e. a countable subset A such that for every $x \in X$ and $\epsilon > 0$ there exists $a \in A$ with $\rho(x,a) < \epsilon$. If $\rho(x,y) = 1 \Leftrightarrow x \neq y$ then ρ is called the *discrete metric*.

The *diameter* of a set $T \subseteq S$ is $\sup\{\rho(x,y) : x, y \in T\}$. If the diameter of T is finite the we say that T is *bounded*. For any $\epsilon > 0$, an *ϵ-cover* for T is a finite set $N \subseteq S$ (not necessarily contained in T) such that for all $x \in T$ there is a $y \in N$ with $\rho(x,y) \leq \epsilon$. If T has a (finite) ϵ-cover for all $\epsilon > 0$ then T is *totally bounded*. (Note that this implies that (T,ρ) is separable and bounded.) In this case the function $\mathcal{N}(\epsilon, T, \rho)$ denotes the size of the smallest ϵ-cover for T (w.r.t. the space S and the (pseudo) metric ρ). We refer to $\mathcal{N}(\epsilon, T, \rho)$ as a *covering number*. A set $R \subseteq T$ is *ϵ-separated* if for all distinct $x, y \in R$, $\rho(x,y) > \epsilon$. We denote by $\mathcal{M}(\epsilon, T, \rho)$ the size of the largest ϵ-separated subset of T. We

computes a weighted sum, so for this unit the modifier is the 0 function and the squashing function is the identity. Since this is the only unit at depth 2 and above, the only Lipschitz bound that counts is the bound for this unit. If β is a bound on the maximum weight coming into the output unit, and W_{max} is the number of units in the hidden layer, then the term $d \log(s(\beta W_{max} + r))$ in the above bound becomes $\log(\beta W_{max})$. Again, fixing $\beta, c_2 - c_1$, M, and c_l gives the same sample size bound,

$$O\left(\frac{1}{\alpha^2 \nu}\left(W \log \frac{W_{max}}{\alpha \nu} + \log \frac{1}{\delta}\right)\right),$$

similar to that obtained in [12].

Since W appears to be the dominant factor in these bounds, apart from the accuracy parameters α and ν, these bounds support the conventional wisdom that the training set size should be primarily related to the number of adjustable parameters in the net. They also support the notion that this relationship between appropriate training size and the number of parameters is nearly linear. Further work is needed to sharpen these relationships (see e.g. the lower bounds obtained in [12]).

8 Conclusion

We have extended the PAC learning model to a more general decision theoretic framework so that it addresses many of the concerns raised by machine learning practitioners, and also introduced a number of new theoretical tools. In this paper we concentrate on applications of the extended model to the problem of obtaining upper bounds on sufficient training sample size. In future papers we hope to deal with the question of lower bounds on sample size needed, and with the computational complexity of finding hypotheses with near optimal observed rate of error. Some promising results along these lines are given in [38]. However, even granting that such results can be obtained, the extended model still has a number of shortcomings in its present form. Some of these can be easily remedied, others may be more problematic.

First, we define the model only for a fixed hypothesis space $\mathcal{H}$. The model should be extended to learning problems on a sequence of hypothesis spaces $\{\mathcal{H}_n : n \geq 1\}$, where $\mathcal{H}_n$ is a hypothesis space on an n-attribute domain X_n (e.g. $[0,1]^n$), and to families of hypothesis spaces of different "complexities" on a fixed domain [37] [16] [34], so that tradeoffs between hypothesis complexity and observed error rate can be addressed. The former extension is easy, the latter more involved. One approach to the latter problem is via Vapnik's principle of structural risk minimization [78] (see also [19]). Other approaches include the MDL (see e.g. [10]) and regularization (see e.g. [59]) methods.

Second, the constants in the upper bounds are still too large to give sample size estimates that are useful in practice. It may be difficult to improve them to the point where the results are directly usable in applied work. Thus even with matching asymptotic lower

computation unit is at most $s(\beta W_{max} + r)$. If this holds for all units at depth 2 and above, may take $b_j = s(\beta W_{max} + r)$ for all $j \geq 2$ in the first bound. Solving for m, this gives the order-of-magnitude estimate of the second bound. □

We give the constants in the upper bound of the above theorem only to show that they are not outlandishly large. We do not mean to suggest that the bound is tight. At present we cannot even verify that the asymptotic bound is tight. In particular, we cannot show that the dependence on the Lipschitz bounds is necessary. Evidence that it may not be necessary comes from the analysis of the case where the squashing function σ is a sharp threshold function, i.e. $\sigma(x) = sign(x)$. Corollary 3 does not apply in this case, because the jump in σ prevents us from obtaining a Lipschitz bound on the computation units. As we let a smooth σ approach the $sign$ function, its slope increases without limit, and thus the bound given in Corollary 3 degenerates. Nevertheless, using the techniques in [12] it can be shown that results similar to Corollary 3 hold in this case, except that no Lipschitz bounds are required, and a bound on the sample size is

$$O\left(\frac{1}{\alpha^2 \nu}\left(W \log \frac{N}{\alpha \nu} + \log \frac{1}{\delta}\right)\right),$$

where N is the total number of computation units in the net. Details are given in Theorem 13 in the Appendix.

Despite these drawbacks, the result does give some indication of the maximum training sample size that will be needed for many popular network configurations. For example, if the squashing function is chosen as $\sigma(x) = 1/(1 + e^{-x/T})$ for some *temperature* $T > 0$ then it can be shown that the Lipschitz bound s for σ is $1/4T$. When the modifier $\mu \equiv 0$, then $r = 0$. Thus in this case the term $d \log(s(\beta W_{max} + r))$ in the bound of Corollary 3 becomes $d \log(\beta W_{max}/T)$. If the maximum weight β, the temperature T and the depth d are constants, along with $c_2 - c_1$, M, and c_l, then the asymptotic bound of the theorem becomes

$$O\left(\frac{1}{\alpha^2 \nu}\left(W \log \frac{W_{max}}{\alpha \nu} + \log \frac{1}{\delta}\right)\right),$$

which is similar to the bound obtained in [12].

Useful bounds are also obtained for networks that implement generalized radial basis[14] functions, as described in [59]. These networks have one layer of hidden units at depth 1 and one output unit at depth 2. The structure of the hidden units is as described in the example above: the input transformers are identity functions, the modifier is $\sum_{i=1}^{n} x_i^2$ and the squashing function is usually a smooth decreasing function. The output unit simply

[14]The computation units in the network of radial basis functions described here are quite primitive in that they have no adjustable multiplicative parameter included in their basic radial distance calculation. Such parameters would be needed to do any reasonable type of kernel based density estimation (see e.g. [20]). These parameters can be simulated by inserting another layer of computation units between the inputs and the layer described here. Alternately, the analysis can also be done directly for adjustable kernel units. This cleaner approach is detailed in [61].

$$\leq \left(\frac{2e(c_2 - c_1)d\prod_{l=2}^{d} b_l}{\epsilon}\right)^{2W}.$$

□

Corollary 3 *Let n, k, $\mathcal{H}$, W, d and $b_2, \ldots, b_d$ be as in the previous theorem. Let X be the instance space $[c_1, c_2]^n$, A be the decision space $[c_1, c_2]^k$, and Y be any outcome space. Let $l : Y \times A \to [0, M]$ be a loss function and c_l be a constant such that $\rho_l(\bar{a}, \bar{b}) \leq c_l d_{L^1}(\bar{a}, \bar{b})$ for all $\bar{a}, \bar{b} \in A$. Let $m \geq 1$, $0 < \nu \leq 8(c_2 - c_1)$, $0 < \alpha < 1$, and P be any probability distribution on $S = X \times Y$. Let $\bar{\xi}$ be generated by m independent random draws from S according to P.*

1. *Then*

$$\mathbf{Pr}\left(\exists f \in \mathcal{H} : d_\nu(\widehat{\mathbf{er}}_{\bar{\xi},l}(f), \mathbf{er}_{P,l}(f)) > \alpha\right) \leq 4\left(\frac{c_l 16e(c_2 - c_1)d\prod_{l=2}^{d} b_l}{\alpha\nu}\right)^{2W} e^{-\alpha^2\nu m/8M}.$$

2. *Assume that for each computation unit at depth 2 and above the number of weights is at most W_{max}, no weight is allowed to have absolute value greater than β, the input transformers are identity functions, the global modifier has Lipschitz bound at most r, and the squashing function has Lipschitz bound at most s, where $s(\beta W_{max} + r) \geq 1$. Then for any $0 < \delta < 1$, the probability in (1.) is less than δ for sample size*

$$m = O\left(\frac{M}{\alpha^2\nu}\left(W\left(\log\frac{c_l(c_2 - c_1)}{\alpha\nu} + d\log(s(\beta W_{max} + r))\right) + \log\frac{1}{\delta}\right)\right).$$

Proof. Let $\epsilon = \alpha\nu/8 \leq c_2 - c_1$. It can be verified that $l_{\mathcal{H}}$ is permissible for the hypothesis space $\mathcal{H}$. Hence, using Theorem 8 and Theorem 10,

$$\begin{aligned}
\mathbf{Pr}\left(\exists f \in \mathcal{H} : d_\nu(\widehat{\mathbf{er}}_{\bar{\xi},l}(f), \mathbf{er}_{P,l}(f)) > \alpha\right) &\leq 4\mathcal{C}(\alpha\nu/8, \mathcal{H}, \rho_l)e^{-\alpha^2\nu m/8M} \\
&\leq 4\mathcal{C}(\alpha\nu/8c_l, \mathcal{H}, d_{L^1})e^{-\alpha^2\nu m/8M} \\
&\leq 4\left(\frac{c_l 16e(c_2 - c_1)d\prod_{l=2}^{d} b_l}{\alpha\nu}\right)^{2W} e^{-\alpha^2\nu m/8M}.
\end{aligned}$$

For the second bound, it is readily verified that the L^1 Lipschitz bound for a linear function defined by W_{max} weights and a bias is no more than W_{max} times the largest absolute value of any weight. Furthermore, the Lipschitz bound for the sum of two functions is no more than the sum of the Lipschitz bounds on the individual functions, and the Lipschitz bound for the composition of two functions is no more than the product of the Lipschitz bounds on the individual functions. Thus if the input transformers are identity functions, the global modifier and squashing function have Lipschitz bounds r and s respectively and no weight is allowed to have absolute value greater than β, then the Lipschitz bound for a

in this manner, by varying the weights and biases in the units $u_1, \ldots, u_{n_j}$ at depth j in such a manner that the Lipschitz constraint is satisfied.

When $j = d$, no subsequent calculations will be performed so we no longer need to preserve the activations of shallower units. Hence, we omit the identity function components in each $h_d \in \mathcal{H}_d$. Otherwise the definition of $\mathcal{H}_d$ is the same as that for $\mathcal{H}_j$, where $j < d$.

It is clear that the class $\mathcal{H}$ in the statement of the theorem can be represented as the class of compositions of functions from classes $\mathcal{H}_1, \ldots, \mathcal{H}_d$. Since the identity function has Lipshitz bound $1 \leq b_j$, the average Lipschitz bound on the components of each function $h_j \in \mathcal{H}_j$ is at most b_j. It is easily verified that a free product function is Lipschitz bounded by the average of the Lipschitz bounds on its component functions. Hence by assumption, b_j is a uniform Lipschitz bound on $\mathcal{H}_j$, $2 \leq j \leq d$. For each j, $1 \leq j \leq d$, let $a_j = \prod_{l=j+1}^{d} b_l$ and $\epsilon_j = \frac{\epsilon}{da_j}$. Since $\epsilon < c_2 - c_1$ and $a_j \geq 1$, $\epsilon_j \leq c_2 - c_1$. Let ρ_j be the d_{L^1} metric on X_j, $1 \leq j \leq d+1$. Then by Lemma 8, part (1),

$$\mathcal{C}(\epsilon, \mathcal{H}, d_{L^1}) \leq \prod_{j=1}^{d} \mathcal{C}(\epsilon_j, \mathcal{H}_j, \rho_{j+1}).$$

For each j, $\mathcal{H}_j$ is contained in the free product of l_j function classes. Each class $\mathcal{F}$ in this product is either the trivial class containing only the identity function, or is a finite dimensional vector space of real-valued functions, summed with a fixed modifier and then composed with a non-increasing or non-decreasing squashing function. In the latter case, the dimension N of this vector space is the number of free parameters associated with the corresponding computation unit, i.e. the number of weights plus one (for the adjustable bias). Hence by Theorems 3 and 4, $\mathbf{dim_P}(\mathcal{F}) \leq N$. Thus by Theorem 9 above,

$$\mathcal{C}(\epsilon_j, \mathcal{F}, d_{L^1}) \leq 2\left(\frac{2e(c_2 - c_1)}{\epsilon_j} \ln \frac{2e(c_2 - c_1)}{\epsilon_j}\right)^N \leq \left(\frac{2e(c_2 - c_1)}{\epsilon_j}\right)^{2N},$$

since $2 \ln x < x$ and $N \geq 1$. Since the capacity of a class with only one function is 1, it follows from Lemma 7 part (1) that

$$\mathcal{C}(\epsilon_j, \mathcal{H}_j, \rho_{j+1}) \leq \left(\frac{2e(c_2 - c_1)}{\epsilon_j}\right)^{2W_j},$$

where W_j is the total number of weights and biases of all computation nodes at depth j. Multiplying these bounds over all j, it follows that

$$\begin{aligned} \mathcal{C}(\epsilon, \mathcal{H}, d_{L^1}) &\leq \prod_{j=1}^{d} \left(\frac{2e(c_2 - c_1)}{\epsilon_j}\right)^{2W_j} \\ &= \prod_{j=1}^{d} \left(\frac{2e(c_2 - c_1)d\prod_{l=j+1}^{d} b_l}{\epsilon}\right)^{2W_j} \end{aligned}$$

function σ to σ', where $\sigma'(x) = \sigma(e^x)$. Then

$$f(\bar{x}) = \sigma' \left(\theta + \sum_{j=1}^{n} w_j \log x_j \right) = \sigma \left(e^{\theta} \prod_{j=1}^{n} x_j^{w_j} \right),$$

giving what is commonly known as a *product unit* [24].

We define a feedforward *architecture* as a feedforward net with unspecified weights and biases, i.e. each computation unit has a fixed global modifier, a fixed squashing function and fixed input transformers, but it has variable weights and a variable bias. We say a unit is *at depth* j in an architecture if the longest (directed) path from an input unit to that unit has j edges. Thus all input units are at depth 0, all computation units that have incoming edges only from input units are at depth 1, all computation units that have incoming edges only from input units and computation units at depth 1 are at depth 2, etc. The *depth of the architecture* is the depth of the deepest unit in it.

We can bound the capacity of the hypothesis space represented by a feedforward architecture as follows.

Theorem 10 *Let $\mathcal{A}$ be a feedforward architecture as above with $n \geq 1$ input units, $k \geq 1$ output units, and depth $d \geq 1$. Let W be the total number of adjustable weights and biases in $\mathcal{A}$. Assume $b_j \geq 1$ for $2 \leq j \leq d$, and let $\mathcal{H}$ be all functions from $[c_0, c_1]^n$ into $[c_0, c_1]^k$ representable on $\mathcal{A}$ by setting the adjustable weights and biases such that for all j, $2 \leq j \leq d$, the average of the Lipschitz bounds of the functions computed by computation units at depth j is at most b_j. Then for all $0 < \epsilon \leq c_2 - c_1$,*

$$\mathcal{C}(\epsilon, \mathcal{H}, d_{L^1}) \leq \left(\frac{2e(c_2 - c_1)d \prod_{l=2}^{d} b_l}{\epsilon} \right)^{2W}.$$

□

Proof. For each j, $0 \leq j \leq d$, let n_j be the number of units at depth j in the architecture $\mathcal{A}$. For each j, $0 \leq j \leq d-1$, let $l_j = \sum_{i=0}^{j} n_j$, and let $l_d = n_d = k$. For each j, $1 \leq j \leq d+1$, let $X_j = [c_1, c_2]^{l_{j-1}}$. Then for each j, $1 \leq j \leq d$, we can define the family $\mathcal{H}_j$ of functions from X_j into X_{j+1} in the following manner.

First assume $j < d$. Let $u_1, \ldots, u_{n_j}$ be an enumeration of the computation units at depth j and $f_1, \ldots, f_{n_j}$ be functions such that f_i can be represented by u_i, $1 \leq i \leq n_j$, and the average Lipschitz bound on the f_is is at most b_j. Let h_j be the free product of $f_1, \ldots, f_{n_j}$ and l_{j-1} copies of the identity function on $[c_1, c_2]$. Thus $h_j : X_j \to X_{j+1}$. The function h_j represents a mapping from the sequence of all activations of units at depth at most $j-1$ to the sequence of all activations of units at depth at most j, where the activations at depth at most $j-1$ are unaltered, and the new activations, i.e. those at depth j, are calculated by $f_1, \ldots, f_{n_j}$. The family $\mathcal{H}_j$ consists of all functions h_j obtained

are not output units are called *hidden units.* The network as a whole computes a function that maps from vectors of activation values in its input units to vectors of activation values in its output units by composing the functions computed by its computation units in the obvious way.

The action of a computation unit with n incoming edges can be specified by a function f from $[c_0, c_1]^n$ into $[c_0, c_1]$, where $f(\bar{x})$ is the resulting activation of the unit when the activations of its incoming edges are given be the vector $\bar{x} \in [c_0, c_1]^n$. In the nets we consider, the function f is defined by

$$f(\bar{x}) = \sigma\left(\mu(\phi_1(\bar{x}), \ldots, \phi_k(\bar{x})) + \theta + \sum_{j=1}^{k} w_j \phi_j(\bar{x})\right),$$

where the w_j's are adjustable real *weights*, θ is an adjustable real *bias*, $\phi_1, \ldots, \phi_k$ are fixed real-valued functions which we call the *input transformers*, $\mu : \Re^k \to \Re$ is a fixed function which we call the *global modifier*, and $\sigma : \Re \to [c_0, c_1]$ is a fixed non-increasing or non-decreasing function which we call the *squashing function.* Different units can have different modifiers, transformers and squashing functions. We say that the function f computed by a given computation unit with n incoming edges has *Lipschitz bound* b if for any $\bar{x}, \bar{y} \in [c_0, c_1]^n$, $|f(\bar{x}) - f(\bar{y})| \leq b d_{L^1}(\bar{x}, \bar{y})$.

We give a few examples to illustrate the flexibility of this model at the level of the individual computation unit. First assume that $k = n$, i.e. the number of input transformers is the same as the number of inputs, and that each input transformer simply extracts a component of the input, i.e. $\phi_j(\bar{x}) = x_j$, $1 \leq j \leq n$. In this case, which is the standard case for most neural net research, the overall input transformation is just the identity map and can be ignored. In this standard case, if the global modifier $\mu = 0$ we get what is known as a quasi-linear unit [66]:

$$f(\bar{x}) = \sigma\left(\theta + \sum_{j=1}^{k} w_j x_j\right).$$

In the standard case, if $\mu(\bar{x}) = \sum_{j=1}^{n} x_j^2$ we get a unit that computes a function of the form

$$f(\bar{x}) = \sigma\left(\theta' + \sum_{j=1}^{n} (x_j - z_j)^2\right),$$

where $z_j = -w_j/2$ and $\theta' = \theta - \sum_{j=1}^{n} z_j^2$. This is similar to what is called a *radial basis* unit in the neural net literature [59,49].

Now assume that $k = n$ but the input transformers take logs of the components of the inputs, i.e. $\phi_j(\bar{x}) = \log x_j$. (Here we assume $c_0 > 0$.) Let $\mu = 0$ and change the squashing

The latter inequality follows from the fact that for $x, y > 0$,

$$|\ln \frac{x}{y}| = \ln \frac{\max(x,y)}{\min(x,y)} \leq \frac{\max(x,y)}{\min(x,y)} - 1 = \frac{|x-y|}{\min(x,y)}.$$

Thus in this case we may take $c_l = 2k/B$.

We now turn to the task of obtaining an upper bound on the capacity $\mathcal{C}(\epsilon, \mathcal{H}, d_{L^1})$. When the decision space $A \subset \Re$, the hypothesis space $\mathcal{H}$ is a family of real valued functions and $d_{L^1}(a,b) = |a-b|$ for $a, b \in A$. In this case we can apply the methods and results of section 4. We must first find an upper bound on $\mathbf{dim_P}(\mathcal{H})$, the pseudo dimension of $\mathcal{H}$. Then, when A is bounded, from the bound on $\mathbf{dim_P}(\mathcal{H})$ we get a bound on the capacity $\mathcal{C}(\epsilon, \mathcal{H}, d_{L^1})$ using Theorem 5. This also gives a bound on the metric dimension of $\mathcal{H}$.

Theorem 9 *Let $\mathcal{H}$ be a family of functions from X into $A = [0, M]$. Assume $\mathbf{dim_P}(\mathcal{H}) = d$ for some $1 \leq d < \infty$.*

1. *For all $0 < \epsilon \leq M$,*
$$\mathcal{C}(\epsilon, \mathcal{H}, d_{L^1}) < 2\left(\frac{2eM}{\epsilon} \ln \frac{2eM}{\epsilon}\right)^d.$$
2. $\overline{\mathbf{dim}}(\mathcal{H}) \leq \mathbf{dim_P}(\mathcal{H})$.

Proof. Let P be any probability measure on X. Then by Theorems 11 and 5

$$\mathcal{N}(\epsilon, \mathcal{H}, d_{L^1(P)}) \leq \mathcal{M}(\epsilon, \mathcal{H}, d_{L^1(P)}) < 2\left(\frac{2eM}{\epsilon} \ln \frac{2eM}{\epsilon}\right)^d.$$

(Theorem 5 is applied with $S = X$ and $\mathbf{F} = \mathcal{H}$.) This gives (1), and (2) follows easily from (1). □

In the general case, where $A \subset \Re^k$ for $k > 1$, we can apply the methods from the previous section, in addition to the methods from section 4, to obtain bounds on $\mathcal{C}(\epsilon, \mathcal{H}, d_{L^1})$. We illustrate this for the case when $\mathcal{H}$ is the class of hypotheses represented by a feedforward neural network.

A feedforward neural network is defined as a directed acyclic graph in which the incoming edges to each node (or *unit*) are ordered and each incoming edge can carry a real number representing the *activation* on that edge. We will assume that all activations are restricted to the interval $[c_0, c_1]$ for some constants $c_0 < c_1$. The units are divided into *input units*, which have no incoming edges from other units and serve as input ports for the network (their activations are determined by these external inputs), and *computation units*, which have incoming edges from other units and compute an activation based on the activations on these incoming edges. After an activation has been determined, this activation is placed on the outgoing edges of the unit. Computation units with no outgoing edges are called *output units* and serve as output ports for the network. Computation units that

For many loss functions, the metric ρ_l on $A \subset \Re^k$ can be bounded in terms of the d_{L^1} metric, i.e. we can find a constant c_l such that for all $\bar{a} = (a_1, \ldots, a_k)$ and $\bar{b} = (b_1, \ldots, b_k)$ in A, $\rho_l(\bar{a}, \bar{b}) \leq c_l d_{L^1}(\bar{a}, \bar{b}) = \frac{c_l}{k} \sum_{i=1}^k |a_i - b_i|$. In this case it is clear that $\mathcal{C}(\epsilon, \mathcal{H}, \rho_l) \leq \mathcal{C}(\epsilon/c_l, \mathcal{H}, d_{L^1})$. Thus our problem is reduced to obtaining an upper bound on the capacity $\mathcal{C}(\epsilon/c_l, \mathcal{H}, d_{L^1})$.

We now give a few examples to illustrate this reduction. First consider the common case in which the outcome space Y is also contained in $\Re^k$, e.g. we receive explicit feedback on each coordinate of our action $\bar{a} \in A \subset \Re^k$. This occurs when each coordinate a_i of the action $\bar{a}$ is a prediction of the corresponding coordinate of the outcome $\bar{y}$. Here the loss function l may itself be a metric on $\Re^k$ which measures the distance between the predicted vector and the actual outcome vector. When l is a metric, we have for any actions $\bar{a}, \bar{b} \in A$

$$\rho_l(\bar{a}, \bar{b}) = \sup_{\bar{y} \in Y} |l(\bar{y}, \bar{a}) - l(\bar{y}, \bar{b})| \leq l(\bar{a}, \bar{b})$$

by the triangle inequality for l. Thus if the metric l is bounded with respect to d_{L^1} metric, i.e. $l(\bar{a}, \bar{b}) \leq c_l d_{L^1}(\bar{a}, \bar{b})$ for all $\bar{a}, \bar{b} \in A$, then we have $\rho_l(\bar{a}, \bar{b}) \leq c_l d_{L^1}(\bar{a}, \bar{b})$. For example, if $l(\bar{y}, \bar{a}) = d_{L^2}(\bar{y}, \bar{a}) = \frac{1}{k} \left(\sum_{i=1}^k (y_i - a_i)^2 \right)^{1/2}$ then we may take $c_l = 1$, and similarly for the other d_{L^q} metrics, for $q > 1$.

Note that the above trick does not apply to the mean squared loss $l(\bar{y}, \bar{a}) = \frac{1}{k} \sum_{i=1}^k (y_i - a_i)^2$ since this loss does not satisfy the triangle inequality. However, in this case it is easy to show by direct calculation that if the outcome space Y is bounded, e.g. $Y \subset [0, M]^k$, then $\rho_l(\bar{a}, \bar{b}) \leq 2M d_{L^1}(\bar{a}, \bar{b})$, and hence we may take $c_l = 2M$.

For our final example, consider the case when $Y = \{0, 1\}^k$, $A \subset [0, 1]^k$ and l is the cross entropy loss

$$l(\bar{y}, \bar{a}) = -\sum_{i=1}^k \left(y_i \ln a_i + (1 - y_i) \ln(1 - a_i) \right).$$

As discussed in section 1.1.3, this is the log likelihood loss for the regression problem in which the action $\bar{a}$ represents a vector of probabilities for independent Bernoulli variables, and the outcome $\bar{y}$ gives the observed values of these variables. This loss is bounded if we restrict the probabilities in $\bar{a}$ to be between B and $1 - B$ for some $0 < B \leq 1/2$. In this case

$$\begin{aligned} \rho_l(\bar{a}, \bar{b}) &= sup_{\bar{y} \in Y} \left| \sum_{i=1}^k \left(y_i \ln \frac{b_i}{a_i} + (1 - y_i) \ln \frac{(1 - b_i)}{(1 - a_i)} \right) \right| \\ &\leq \sum_{i=1}^k |\ln \frac{b_i}{a_i}| + \sum_{i=1}^k |\ln \frac{(1 - b_i)}{(1 - a_i)}| \\ &\leq \frac{2}{B} \sum_{i=1}^k |a_i - b_i|. \end{aligned}$$

$$
\begin{aligned}
& + \int_{X_1} \rho_{h+1}(f_h \circ f_{h-1} \circ \cdots \circ f_1(x), g_h \circ f_{h-1} \circ \cdots \circ f_1(x)) dP(x) \\
\leq\ & b_h \int_{X_1} \rho_h(f_{h-1} \circ \cdots \circ f_1(x), g_{h-1} \circ \cdots \circ g_1(x)) dP(x) \\
& + \int_{X_h} \rho_{h+1}(f_h(y), g_h(y)) dP_{f_{h-1},\ldots,f_1}(y) \quad (by\ Lip.\ assump.) \\
\leq\ & b_h \sum_{j=1}^{h-1} \left(\prod_{l=j+1}^{h-1} b_l \right) \epsilon_j + \epsilon_h \quad (by\ i.h.\ and\ def.\ of\ f_h) \\
=\ & \sum_{j=1}^{h} \left(\prod_{l=j+1}^{h} b_l \right) \epsilon_j .
\end{aligned}
$$

To prove part (2), let $a_j = \prod_{l=j+1}^{k} b_l$ and set $\epsilon_j = \frac{\epsilon}{k a_j}$ for $1 \leq j \leq k$. By part (1), $\mathcal{C}(\epsilon, \mathcal{H}, \rho_{k+1}) \leq \prod_{j=1}^{k} \mathcal{C}(\epsilon_j, \mathcal{H}_j, \rho_{j+1})$. Thus

$$
\begin{aligned}
\overline{\mathbf{dim}}(\mathcal{H}) &= \mathbf{limsup}_{\epsilon \to 0} \left(\frac{\log(\mathcal{C}(\epsilon, \mathcal{H}, \rho_{k+1}))}{\log(1/\epsilon)} \right) \\
&\leq \mathbf{limsup}_{\epsilon \to 0} \left(\frac{\log(\prod_{j=1}^{k} \mathcal{C}(\epsilon/k a_j, \mathcal{H}_j, \rho_{j+1}))}{\log(1/\epsilon)} \right) \\
&= \sum_{j=1}^{k} \mathbf{limsup}_{\epsilon \to 0} \left(\frac{\log(\mathcal{C}(\epsilon/k a_j, \mathcal{H}_j, \rho_{j+1}))}{\log(1/\epsilon)} \right) \\
&= \sum_{j=1}^{k} \mathbf{limsup}_{\epsilon_j \to 0} \left(\frac{\log(\mathcal{C}(\epsilon_j, \mathcal{H}_j, \rho_{j+1}))}{\log(1/\epsilon_j) + \log(1/k a_j)} \right) \\
&= \sum_{j=1}^{k} \mathbf{limsup}_{\epsilon_j \to 0} \left(\frac{\log(\mathcal{C}(\epsilon_j, \mathcal{H}_j, \rho_{j+1}))}{\log(1/\epsilon_j)} \right) \\
&= \sum_{j=1}^{k} \overline{\mathbf{dim}}(\mathcal{H}_j).
\end{aligned}
$$

□

7 Sample size bounds for learning with multi-layer neural nets

We now present some applications of the results of the previous section to learning with feedforward neural nets (see e.g. [66,59]). The hypothesis space $\mathcal{H}$ represented by a feedforward neural net consists of a family of functions from an instance space $X \subset \Re^n$ into a decision space $A \subset \Re^k$ for some $k, n \geq 1$. To apply Theorem 8 of the previous section, we will need to obtain an upper bound on the capacity $\mathcal{C}(\epsilon, \mathcal{H}, \rho_l)$ of such hypothesis spaces for various loss functions l.

2. $\overline{\mathbf{dim}}(\mathcal{H}) \leq \sum_{j=1}^{k} \overline{\mathbf{dim}}(\mathcal{H}_j)$*, and similarly for* $\underline{\mathbf{dim}}$ *and* $\mathbf{dim}$*, when the latter is defined.*

Proof. Fix a probability measure P on X_1. We define a tree-structured family of covers for the $\mathcal{H}_j$'s by induction as follows. For the basis case, let U be a minimum-sized ϵ_1-cover for $\mathcal{H}_1$ w.r.t. the measure P on X_1, i.e. $|U| = \mathcal{N}(\epsilon_1, \mathcal{H}_1, d_{L^1(P,\rho_2)})$ and every function in $\mathcal{H}_1$ is $L^1(P, \rho_2)$-approximated to within ϵ_1 by some function in U. Now for each j, $2 \leq j \leq k$, and for each sequence of functions $f_1, \ldots, f_{j-1}$ where $f_1 \in U$, $f_2 \in U_{f_1}$, $f_3 \in U_{f_1,f_2}$, ..., $f_{j-1} \in U_{f_1,\ldots,f_{j-2}}$, let $U_{f_1,\ldots,f_{j-1}}$ be a minimum-sized ϵ_j cover for $\mathcal{H}_j$ w.r.t. the L^1 metric for the measure $P_{f_{j-1} \circ f_{j-2} \circ \cdots \circ f_1}$ on X_j and the metric ρ_{j+1} on X_{j+1}.

Next we define a cover V for $\mathcal{H}$ by composing functions in the covers for the $\mathcal{H}_j$'s. If $k = 1$ then $V = U$. Otherwise

$$V = \{f_k \circ f_{k-1} \circ \cdots \circ f_1 : f_1 \in U, f_2 \in U_{f_1}, f_3 \in U_{f_1,f_2}, \cdots, \textit{and } f_k \in U_{f_1,\ldots,f_{k-1}}\}.$$

Since $\mathcal{N}(\epsilon_j, \mathcal{H}_j, d_{L^1(P_{f_{j-1} \circ f_{j-2} \circ \cdots \circ f_1}, \rho_{j+1})}) \leq \mathcal{C}(\epsilon_j, \mathcal{H}_j, \rho_{j+1})$ for all $1 \leq j \leq k$ and all $f_1, \ldots, f_{j-1}$, it is clear that $|V| \leq \prod_{j=1}^{k} \mathcal{C}(\epsilon_j, \mathcal{H}_j, \rho_{j+1})$. Hence it remains to show that V is an ϵ-cover for $\mathcal{H}$.

Suppose that $g = g_k \circ g_{k-1} \circ \cdots \circ g_1 \in \mathcal{H}$. Find

- $f_1 \in U$ such that $d_{L^1(P,\rho_2)}(f_1, g_1) \leq \epsilon_1$,
- $f_2 \in U_{f_1}$ such that $d_{L^1(P_{f_1},\rho_3)}(f_2, g_2) \leq \epsilon_2$, ..., and
- $f_k \in U_{f_1,\ldots,f_{k-1}}$ such that $d_{L^1(P_{f_{k-1} \circ f_{k-2} \circ \cdots \circ f_1}, \rho_{k+1})}(f_k, g_k) \leq \epsilon_k$.

Let $f = f_k \circ f_{k-1} \circ \cdots \circ f_1 \in V$. It suffices to show that $d_{L^1(P,\rho_{k+1})}(f, g) \leq \epsilon$. We prove that for all h, $1 \leq h \leq k$,

$$d_{L^1(P,\rho_{h+1})}(f_h \circ \cdots \circ f_1, g_h \circ \cdots \circ g_1) \leq \sum_{j=1}^{h} \left(\prod_{l=j+1}^{h} b_l \right) \epsilon_j.$$

Since

$$\epsilon = \sum_{j=1}^{k} \left(\prod_{l=j+1}^{k} b_l \right) \epsilon_j,$$

part (1) of the result follows.

If $h = 1$ then the result follows directly from our definition of f. Otherwise

$$\begin{aligned} & d_{L^1(P,\rho_{h+1})}(f_h \circ \cdots \circ f_1, g_h \circ \cdots \circ g_1) \\ = & \int_{X_1} \rho_{h+1}(f_h \circ \cdots \circ f_1(x), g_h \circ \cdots \circ g_1(x)) dP(x) \\ \leq & \int_{X_1} \rho_{h+1}(g_h \circ f_{h-1} \circ \cdots \circ f_1(x), g_h \circ g_{h-1} \circ \cdots \circ g_1(x)) dP(x) \end{aligned}$$

Hence U is an ϵ-cover for $\mathcal{H}$.

The first inequality of (1) is verified similarly. For each $1 \leq j \leq k$ let V_j be an $k\epsilon$-separated subset of $\mathcal{H}_j$. Let $V = \prod_{j=1}^{k} V_j$. Let $f = (f_1, \ldots, f_k)$ and $g = (g_1, \ldots, g_k)$ be distinct functions in V. Then

$$d_{L^1(P,\rho)}(f,g) = \frac{1}{k}\sum_{j=1}^{k} d_{L^1(P,\rho_j)}(f_j, g_j) > \epsilon.$$

Hence V is an ϵ-separated subset of $\mathcal{H}$. It follows that $\prod_{j=1}^{k} \mathcal{M}(k\epsilon, \mathcal{H}_j, d_{L^1(P,\rho_j)}) \leq \mathcal{M}(\epsilon, \mathcal{H}, d_{L^1(P,\rho)})$. The first inequality of (1) then follows using Theorem 11.

From (1) we have

$$\prod_{j=1}^{k} \mathcal{C}(2k\epsilon, \mathcal{H}_j, \rho_j) \leq \mathcal{C}(\epsilon, \mathcal{H}, \rho) \leq \prod_{j=1}^{k} \mathcal{C}(\epsilon, \mathcal{H}_j, \rho_j).$$

Part (2) follows easily from this. □

Definition 9 *Let P be a probability measure on X and f be a measurable function from X into Y. Then P_f denotes the probability measure on Y induced by f, i.e.*

$$P_f(S) = P(f^{-1}(S)) \text{ for all measurable } S \subset Y.$$

Definition 10 *Let f be a function from a metric space (X, ρ) into a metric space (Y, σ). A* Lipschitz bound *on f is a real number $b > 0$ such that for all $x, y \in X$, $\sigma(f(x), f(y)) \leq b\,\rho(x, y)$.* The *Lipschitz bound on f is the smallest such b. If $\mathcal{F}$ is a class of functions from (X, ρ) into (Y, σ) then b is a* uniform Lipschitz bound *on $\mathcal{F}$ if b is a Lipschitz bound on f for all $f \in \mathcal{F}$.*

Lemma 8 *Let $(X_1, \rho_1), \ldots, (X_{k+1}, \rho_{k+1})$ be metric spaces, where (X_j, ρ_j) is bounded, $2 \leq j \leq k$, and $\mathcal{H}_j$ be a class of functions with $f : X_j \to X_{j+1}$ for all $f \in \mathcal{H}_j$, $1 \leq j \leq k$. Let b_j be a uniform Lipschitz bound on $\mathcal{H}_j$ for all $2 \leq j \leq k$. Let $\mathcal{H}$ denote the class of all functions from X_1 into X_{k+1} defined by compositions of functions in the $\mathcal{H}_j$'s, i.e.*

$$\mathcal{H} = \{f_k \circ f_{k-1} \circ \cdots \circ f_1 : f_j \in \mathcal{H}_j, 1 \leq j \leq k\}.$$

1. *For any $\epsilon, \epsilon_1, \ldots, \epsilon_k > 0$ such that*

$$\epsilon = \sum_{j=1}^{k} \left(\prod_{l=j+1}^{k} b_l \right) \epsilon_j$$

we have

$$\mathcal{C}(\epsilon, \mathcal{H}, \rho_{k+1}) \leq \prod_{j=1}^{k} \mathcal{C}(\epsilon_j, \mathcal{H}_j, \rho_{j+1}).$$

In order to apply the above theorem, we need tools for bounding the capacity of various hypothesis spaces. Along these lines, we close this section by proving two basic lemmas, one about the capacity of the free product of a set of function classes, and the other about the capacity of compositions of functions classes.

Definition 8 *Let $(A_1, \rho_1), \ldots, (A_k, \rho_k)$ be bounded metric spaces. Let $A = A_1 \times \cdots \times A_k$ and ρ be the metric on A defined by*

$$\rho(\bar{u}, \bar{v}) = \frac{1}{k} \sum_{j=1}^{k} \rho_j(u_j, v_j)$$

for any $\bar{u} = (u_1, \ldots, u_k)$ and $\bar{v} = (v_1, \ldots, v_k) \in A$. For each j, $1 \leq j \leq k$, let $\mathcal{H}_j$ be a family of functions from X into A_j. The free product *of $\mathcal{H}_1$ through $\mathcal{H}_k$ is the class of functions*

$$\mathcal{H} = \{(f_1, \ldots, f_k) : f_j \in \mathcal{H}_j, 1 \leq j \leq k\},$$

where $(f_1, \ldots, f_k) : X \to A$ is the function defined by

$$(f_1, \ldots, f_k)(x) = (f_1(x), \ldots, f_k(x)).$$

Lemma 7 *If $\mathcal{H}, \mathcal{H}_1, \ldots, \mathcal{H}_k$ are defined as above then*

1. *for any probability measure P on X and $\epsilon > 0$,*

$$\prod_{j=1}^{k} \mathcal{N}(2k\epsilon, \mathcal{H}_j, d_{L^1(P,\rho_j)}) \leq \mathcal{N}(\epsilon, \mathcal{H}, d_{L^1(P,\rho)}) \leq \prod_{j=1}^{k} \mathcal{N}(\epsilon, \mathcal{H}_j, d_{L^1(P,\rho_j)}),$$

2. *$\overline{\mathbf{dim}}(\mathcal{H}) = \sum_{j=1}^{k} \overline{\mathbf{dim}}(\mathcal{H}_j)$, and similarly for $\underline{\mathbf{dim}}$ and $\mathbf{dim}$, when the latter is defined.*

Proof. We begin with the second inequality of (1). For each $1 \leq j \leq k$ let U_j be an ϵ-cover for $\mathcal{H}_j$. Let

$$U = \prod_{j=1}^{k} U_j = \{(f_1, \ldots, f_k) : f_j \in U_j, 1 \leq j \leq k\}.$$

It suffices to show that U is an ϵ-cover for $\mathcal{H}$. Let $g = (g_1, \ldots, g_k)$ be any function in $\mathcal{H}$. For each j, $1 \leq j \leq k$, find $f_j \in U_j$ such that $d_{L^1(P,\rho_j)}(f_j, g_j) \leq \epsilon$. Let $f = (f_1, \ldots, f_k)$. Then

$$\begin{aligned} d_{L^1(P,\rho)}(f, g) &= \int_X \frac{1}{k} \sum_{j=1}^{k} \rho_j(f_j(x), g_j(x)) dP(x) \\ &= \frac{1}{k} \sum_{j=1}^{k} \int_X \rho_j(f_j(x), g_j(x)) dP(x) \\ &= \frac{1}{k} \sum_{j=1}^{k} d_{L^1(P,\rho_j)}(f_j, g_j) \\ &\leq \epsilon. \end{aligned}$$

We now show how bounds on the capacity of $\mathcal{H}$ lead to distribution-independent bounds on the rate of uniform convergence of empirical error estimates for functions in $\mathcal{H}$ with respect to the loss function l. As before, let $S = X \times Y$, P be a probability distribution on S, and $l_{\mathcal{H}}$ be the family of functions on S defined by $l_{\mathcal{H}} = \{l_h : h \in \mathcal{H}\}$, where $l_h(x, y) = l(y, h(x))$. Let $P_{|X}$ be the marginal on X of the joint distribution P on $X \times Y$ (see section 1.5).

Lemma 6 *For all $\epsilon > 0$,*

$$\mathcal{N}(\epsilon, l_{\mathcal{H}}, d_{L^1(P)}) \leq \mathcal{N}(\epsilon, \mathcal{H}, d_{L^1(P_{|X}, \rho_l)}).$$

Proof. For every $h \in \mathcal{H}$ let $\psi(h) = l_h$. Hence ψ maps from $\mathcal{H}$ onto $l_{\mathcal{H}}$. It suffices to show that ψ is a contraction, i.e. that

$$\forall\, f, g \in \mathcal{H},\ d_{L^1(P)}(\psi(f), \psi(g)) \leq d_{L^1(P_{|X}, \rho_l)}(f, g).$$

Let f and g be any two functions in $\mathcal{H}$. Then

$$\begin{aligned} d_{L^1(P)}(\psi(f), \psi(g)) &= \int_S |l(y, f(x)) - l(y, g(x))| dP(x, y) \\ &\leq \int_S \rho_l(f(x), g(x)) dP(x, y) \\ &= \int_X \rho_l(f(x), g(x)) dP_{|X}(x) \\ &= d_{L^1(P_{|X}, \rho_l)}(f, g). \end{aligned}$$

□

This gives the following theorem about uniform convergence of error estimates for learning.

Theorem 8 *Assume that the hypothesis space $\mathcal{H}$ and the loss function l are such that $l_{\mathcal{H}}$ is permissible. Let P be any probability distribution on $S = X \times Y$. Assume $m \geq 1$, $\nu > 0$ and $0 < \alpha < 1$. Let $\bar{\xi}$ be generated by m independent draws from S according to P. Then*

$$\mathbf{Pr}\left(\exists h \in \mathcal{H} : d_\nu(\widehat{\mathbf{er}}_{\bar{\xi},l}(h), \mathbf{er}_{P,l}(h)) > \alpha\right) \leq 4\mathcal{C}(\alpha\nu/8, \mathcal{H}, \rho_l) e^{-\alpha^2 \nu m / 8M}.$$

Proof. Let $\mathbf{F} = l_{\mathcal{H}}$. For any sequence $\bar{\xi}$ of points in S there is a trivial isometry between $(\mathbf{F}_{|\bar{\xi}}, d_{L^1})$ and $(\mathbf{F}, d_{L^1(P_{\bar{\xi}})})$, where $P_{\bar{\xi}}$ is the empirical distribution induced by $\bar{\xi}$, as in the proof of Theorem 6. Thus by Lemma 6 above, we have

$$\mathcal{N}(\epsilon, \mathbf{F}_{|\bar{\xi}}, d_{L^1}) = \mathcal{N}(\epsilon, \mathbf{F}, d_{L^1(P_{\bar{\xi}})}) \leq \mathcal{N}(\epsilon, \mathcal{H}, d_{L^1(P_{\bar{\xi}|X}, \rho_l)}) \leq \mathcal{C}(\epsilon, \mathcal{H}, \rho_l)$$

for all $\bar{\xi} \in S^{2m}$. Hence, setting $\epsilon = \alpha\nu/8$, the given probability is at most $4\mathcal{C}(\alpha\nu/8, \mathcal{H}, \rho_l) e^{-\alpha^2 \nu m / 8M}$ by Theorem 12 in section 10.4. □

Definition 6 *For every loss function $l : Y \times A \to [0, M]$, by ρ_l we denote the pseudo metric on A defined by $\rho_l(a, b) = \sup_{y \in Y} |l(y, a) - l(y, b)|$ for all $a, b \in A$.*

Note that (A, ρ_l) is a bounded pseudo metric space: no two actions in A are more than M apart.

Since hypotheses in $\mathcal{H}$ map from the instance space X into A, the pseudo metric ρ_l on A can be used to induce a pseudo metric on $\mathcal{H}$ in which two hypotheses differ only to the extent that the actions that they proscribe differ with respect to ρ_l. There are several ways to do this. The easiest is to use an L^∞ function distance on $\mathcal{H}$, defining the distance between hypotheses f and g as the supremum of $\rho_l(f(x), g(x))$ over all $x \in X$. This works, and is a useful method of obtaining uniform convergence and learning results (see related techniques used in [83]). However, as we will see, the crucial issue is the size of the smallest ϵ-cover of the resulting pseudo metric space $\mathcal{H}$. In some cases we can get smaller covers, and hence better results, by using an L^1 function distance instead. Since the L^1 distance is never more than the L^∞ distance, the results are never worse. Thus we present this more powerful method here.

Definition 7 *Let $\mathcal{H}$ be a family of functions from a set X into a bounded pseudo metric space (A, ρ). Let P be a probability measure on X. Then $d_{L^1(P,\rho)}$ is the pseudo metric on $\mathcal{H}$ defined by*

$$d_{L^1(P,\rho)}(f, g) = \mathbf{E}(\rho(f(x), g(x))) = \int_X \rho(f(x), g(x)) dP(x) \text{ for all } f, g \in \mathcal{H}.$$

For every $\epsilon > 0$ let

$$\mathcal{C}(\epsilon, \mathcal{H}, \rho) = sup\{\mathcal{N}(\epsilon, \mathcal{H}, d_{L^1(P,\rho)})\} \text{ over all probability measures } P \text{ on } X.$$

If $\mathcal{N}(\epsilon, \mathcal{H}, d_{L^1(P,\rho)})$ is infinite for some measure P, or if the set in this supremum is unbounded, then $\mathcal{C}(\epsilon, \mathcal{H}, \rho) = \infty$. We call [13] *$\mathcal{C}(\epsilon, \mathcal{H}, \rho)$ the* capacity *of $\mathcal{H}$. In analogy with the definition of metric dimension, we define the* upper metric dimension *of $\mathcal{H}$ by*

$$\overline{\mathbf{dim}}(\mathcal{H}) = \mathbf{limsup}_{\epsilon \to 0} \frac{\log \mathcal{C}(\epsilon, \mathcal{H}, \rho)}{\log(1/\epsilon)},$$

and the lower metric dimension*, denoted by $\underline{\mathbf{dim}}(\mathcal{H})$, is defined similarly using* **liminf**. *When $\overline{\mathbf{dim}}(\mathcal{H}) = \underline{\mathbf{dim}}(\mathcal{H})$, then this quantity is denoted $\mathbf{dim}(\mathcal{H})$, and referred to simply as the* metric dimension *of $\mathcal{H}$. If $\mathcal{C}(\epsilon, \mathcal{H}, \rho) = \infty$ for some $\epsilon > 0$ then $\mathbf{dim}(\mathcal{H}) = \infty$.*

[13] The term *metric entropy* is often used for the quantities $\log \mathcal{N}(\epsilon, \mathcal{H}, d_{L^1(P,\rho)})$ and $\log \mathcal{C}(\epsilon, \mathcal{H}, \rho)$ [23,63]. It is also used for an analogous, but fundamentally distinct, concept in the dynamical systems literature (e.g. [27]). The term *capacity* has also been used with many other related meanings [47,77,40,28,12]. Our usage here is taken from [22].

Further applications, specifically for learning problems that have been studied recently in the mainstream artificial intelligence work, are given in [31,32].

For our second example, consider the case that the outcome space Y is $\{0,1\}$ but the decision space A is $[0,1]$. Assume that the loss function is $l(y,a) = |a-y|^q$ for some $q > 0$. This case was examined in [38] with $q = 2$ by Kearns and Schapire in their investigation into the learnability of p-concepts. They showed that $\mathbf{dim_P}(l_{\mathcal{H}}) = \mathbf{dim_P}(\mathcal{H})$ in this case. Since the loss l is monotone in A for all $q > 0$, Lemma 5 shows that this result holds for other values of q as well. Hence the conditions of Theorem 7 are met. Some applications are given in [38]. (It should be noted that it is important that $A = [0,1]$ in this case. The result does not hold in general for larger A.)

For our third example, consider the problem of logistic regression, as described in section 1.1.3. In the simplest case the outcome space again has only two values, denoted y_1 and y_2, where y_1 indicates that some event has taken place and y_2 indicates that it has not, and an action a represents an estimate of the log odds ratio $\ln(P(y_1)/P(y_2))$, where the probability P is conditioned on the observed instance x. Here $A = \Re$ and the log likelihood loss function is the logistic loss function, defined by $l(y_1,a) = \ln(1+e^a) - a$ and $l(y_2,a) = \ln(1+e^a)$. Again, it is easily verified that l is monotone in A. In logistic regression the standard assumption is that the instance space X is contained in $\Re^n$ for some $n \geq 1$ and $\mathcal{H}$ is contained in the family of all linear functions on X (see e.g. [48]). In this case, $\mathbf{dim_P}(\mathcal{H}) \leq n+1$ by Theorem 3. By restricting A to a bounded range, we can then apply Theorem 7 to obtain sample complexity bounds that are linear in n.

As a last example, consider the problem of density estimation, as described in section 1.1.4. Here there is only one outcome in Y, $A \subset \Re^+$, and $l(y,a) = l(a) = -\log a$. Thus clearly l is monotone in A. Thus we can apply Theorem 7 to the problem of denisity estimation as well, whenever the family of densities $\mathcal{H}$ is uniformly bounded (away from zero) and has finite pseudo dimension.

6 Capacity and Metric Dimension of Function Classes

In sections 4 and 5 we showed how the pseudo dimension can be used to obtain distribution independent bounds on the random covering numbers needed for Theorem 2, thereby obtaining bounds on the sample size needed for uniform convergence and learning results. In this section we develop an alternate way of obtaining distribution independent bounds on random covering numbers. This method can sometimes be used in conjunction with the method given in the previous sections. We will demonstrate this in our analysis of the sample size needed for learning in feedforward neural networks in the following section.

The key idea is to introduce a pseudo metric on the decision space A. The distance between two actions is the maximum difference in loss for these actions, over all possible outcomes.

Moreover, for $m \geq \frac{8M}{\alpha^2\nu}\left(2d\ln\frac{8eM}{\alpha\nu} + \ln\frac{8}{\delta}\right)$ *this probability is at most* δ.

Proof. By Lemma 5, $\mathbf{dim_P}(l_{\mathcal{H}}) = \mathbf{dim_P}(\mathcal{H})$ in this case. Thus, since $\widehat{\mathbf{er}}_{\bar{\xi},l}(h) = \hat{\mathbf{E}}_{\bar{\xi}}(l_h)$ and $\mathbf{er}_{P,l}(h) = \mathbf{E}(l_h)$, the result follows directly from Theorem 6. □

When its conditions are satisfied, this theorem, combined with Lemma 1 from section 2.5, gives us a means of bounding in terms of $\mathbf{dim_P}(\mathcal{H})$ the sample complexity $m(\alpha, \nu, \delta)$ of any algorithm that solves the basic learning problem by returning (with high probability) a hypothesis with near minimal empirical error on the training sample. The resulting bound is

$$m(\alpha, \nu, \delta) = O\left(\frac{M}{\alpha^2\nu}\left(\mathbf{dim_P}(\mathcal{H})\log\frac{M}{\alpha\nu} + \log\frac{1}{\delta}\right)\right). \tag{5}$$

This is similar to the sample complexity

$$m(\alpha, \nu, \delta) = O\left(\frac{M}{\alpha^2\nu}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right)\right)$$

that can be obtained by using Theorem 1, when $\mathcal{H}$ (and hence $l_{\mathcal{H}}$) is finite. The term $\mathbf{dim_P}(\mathcal{H})\log\frac{M}{\alpha\nu}$ replaces the term $\log|\mathcal{H}|$. In particular, for the "PAC settings" $\alpha = 1/2$ and $\nu = \epsilon$ we get the sample complexity

$$m(\epsilon, \delta) = O\left(\frac{M}{\epsilon}\left(\mathbf{dim_P}(\mathcal{H})\log\frac{M}{\epsilon} + \log\frac{1}{\delta}\right)\right), \tag{6}$$

in place of the sample complexity

$$m(\epsilon, \delta) = O\left(\frac{M}{\epsilon}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right)\right)$$

derived from Theorem 1. Moreover, these bounds are distribution-independent, so $\mathcal{P}$ can be taken to be the class of all probability distributions on S.

We now give several examples to illustrate the use of this theorem. First, consider the standard PAC model in which the outcome space Y and the decision space A are both $\{0, 1\}$. In this case any loss function such that $l(0,1) \neq l(0,0)$ and $l(1,1) \neq l(1,0)$ is monotone over A, and in particular, the standard discrete loss function, $l(y, a) = 1$ if $y = a$, else $l(y, a) = 0$, is monotone. Clearly $M = 1$ in this case. Thus from equation (6) above we get a sample complexity bound of

$$O\left(\frac{1}{\epsilon}\left(\mathbf{dim_P}(\mathcal{H})\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right).$$

As mentioned in the previous section, $\mathbf{dim_P}(\mathcal{H})$ is the same as the Vapnik-Chervonenkis dimension of $\mathbf{H}$ in this case, hence this bound is, up to constants, the same as that given in Theorem 2.1 of [16]. A number of applications of this result are outlined in [16].