[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].

Proof. Suppose that $\bar{x} = (x_1, \ldots, x_d)$ is shattered by $\mathcal{H}$, where $x_i \in X$, $1 \leq i \leq d$. Then there is some real vector $\bar{r} = (r_1, \ldots, r_d)$ such that $\mathcal{H}_{|\bar{x}} - \bar{r}$ intersects all 2^d orthants of $\Re^d$. Hence for every Boolean vector $\bar{b} \in \{0,1\}^d$ there exists a function $h_{\bar{b}} \in \mathcal{H}$ such that for every i, $1 \leq i \leq d$, we have $h_{\bar{b}}(x_i) > r_i$ if and only if the i^{th} bit of $\bar{b}$ is 1. Fix an outcome $y \in Y$. Let $\bar{\xi} = (\xi_1, \ldots, \xi_d)$, where $\xi_i = (x_i, y)$ for all i, $1 \leq i \leq d$. Note that if f_y is strictly increasing, then for any $h \in \mathcal{H}$ and $1 \leq i \leq d$, $h(x_i) > r_i \Leftrightarrow l_h(\xi_i) = l(y, h(x_i)) = f_y(h(x_i)) > f_y(r_i)$. Hence, for every Boolean vector $\bar{b} \in \{0,1\}^d$ there exists a function $h_{\bar{b}} \in \mathcal{H}$ such that for every i, $1 \leq i \leq d$, we have $l_{h_{\bar{b}}}(\xi_i) > f_y(r_i)$ if and only if the i^{th} bit of $\bar{b}$ is 1. A similar result holds if f_y is strictly decreasing. Thus $\bar{\xi}$ is shattered by $l_{\mathcal{H}}$. It follows that $\mathbf{dim_P}(l_{\mathcal{H}}) \geq \mathbf{dim_P}(\mathcal{H})$.

For the other direction, assume $\bar{\xi} = (\xi_1, \ldots, \xi_d)$ is shattered by $l_{\mathcal{H}}$, where $\xi_i = (x_i, y_i)$ for all i, $1 \leq i \leq d$. We will show that $\bar{x} = (x_1, \ldots, x_d)$ is shattered by $\mathcal{H}$. Since $\bar{\xi}$ is shattered by $l_{\mathcal{H}}$, there is some real vector $\bar{r} = (r_1, \ldots, r_d)$ such that for every Boolean vector $\bar{b} \in \{0,1\}^d$ there exists a function $h_{\bar{b}} \in \mathcal{H}$ such that for every i, $1 \leq i \leq d$, we have $l_{h_{\bar{b}}}(\xi_i) > r_i$ if and only if the i^{th} bit of $\bar{b}$ is 1. Let $A_0 = \{h_{\bar{b}}(x_i) : 1 \leq i \leq d \text{ and } \bar{b} \in \{0,1\}^d\}$. For each outcome $y \in Y$ define the function $g_y : \Re \to A_0$ as follows. If f_y is increasing then $g_y(r) = max\{a \in A_0 : f_y(a) \leq r\}$ and if f_y is decreasing then $g_y(r) = max\{a \in A_0 : f_y(a) > r\}$. Then for each i, $1 \leq i \leq d$, we either have

1. for all $h \in \mathcal{H}$, $l_h(\xi_i) = f_{y_i}(h(x_i)) > r_i \Leftrightarrow h(x_i) > g_{y_i}(r_i)$ or
2. for all $h \in \mathcal{H}$, $l_h(\xi_i) = f_{y_i}(h(x_i)) > r_i \Leftrightarrow h(x_i) \leq g_{y_i}(r_i)$

Hence, for every Boolean vector $\bar{b} \in \{0,1\}^d$ there exists a function $h'_{\bar{b}} \in \mathcal{H}$ such that for every i, $1 \leq i \leq d$, we have $h'_{\bar{b}}(x_i) > g_{y_i}(r_i)$ if and only if the i^{th} bit of $\bar{b}$ is 1. (To see this, let $\bar{c}$ be the Boolean vector derived from $\bar{b}$ by complementing the bit in each position i for which f_{y_i} is decreasing, and then let $h'_{\bar{b}} = h_{\bar{c}}$.) Thus $\bar{x} = (x_1, \ldots, x_d)$ is shattered by $\mathcal{H}$. It follows that $\mathbf{dim_P}(\mathcal{H}) \geq \mathbf{dim_P}(l_{\mathcal{H}})$, and combined with the above inequality, this gives the result. □

Combined with Theorem 6, this gives the following result on the uniform convergence of empirical error estimates for the basic learning problem. Here and below we use the notation $\widehat{\mathbf{er}}_{\bar{\xi},l}(h)$ and $\mathbf{er}_{P,l}(h)$ introduced in section 2 for the empirical error estimate and true error of a hypothesis h, respectively.

Theorem 7 *Assume the decision space $A \subset \Re$, the loss function l is monotone over A and bounded between 0 and M, the hypothesis space $\mathcal{H}$ is such that $l_{\mathcal{H}}$ is permissible, and $1 \leq d = \mathbf{dim_P}(\mathcal{H}) < \infty$. Assume $m \geq 1$, $0 < \nu < 8M$ and $0 < \alpha < 1$. Let P be any probability distribution on S. Let $\bar{\xi}$ be generated by m independent draws from S according to P. Then*

$$\mathbf{Pr}\left(\exists h \in \mathcal{H} : d_\nu(\widehat{\mathbf{er}}_{\bar{\xi},l}(h), \mathbf{er}_{P,l}(h)) > \alpha\right) \leq 8\left(\frac{16eM}{\alpha\nu}\ln\frac{16eM}{\alpha\nu}\right)^d e^{-\alpha^2\nu m/8M}.$$

Corollary 2 *Under the same assumptions as above, for all* $0 < \epsilon \leq M$,

$$\mathbf{Pr}\left(\exists f \in \mathbf{F} : |\widehat{\mathbf{E}}_{\bar{\xi}}(f) - \mathbf{E}(f)| > \epsilon\right) \leq 8\left(\frac{32eM}{\epsilon}\ln\frac{32eM}{\epsilon}\right)^{d} e^{-\epsilon^2 m/64M^2}.$$

Moreover, for $m \geq \frac{64M^2}{\epsilon^2}\left(2d\ln\frac{16eM}{\epsilon} + \ln\frac{8}{\delta}\right)$ *this probability is at most* δ.

Proof. This follows directly from the above result by setting $\nu = 2M$, $\alpha = \epsilon/4M$, and using property (3) of the d_ν metric, as in the proof of Corollary 1 in section 3.2. □

5 Some Applications of Pseudo Dimension in Learning

We now look at how the theoretical results obtained in the previous two sections can be applied to certain types of learning problems. Suppose that we have a basic learning problem defined by X, Y, A, $\mathcal{H}$, $\mathcal{P}$, and $\mathcal{L}$, where $\mathcal{L}$ is the family of $L_{\alpha,\nu}$ regret functions for an underlying loss function l, as in section 2. As before, let $S = X \times Y$, $l_h : S \to [0, M]$ be defined by $l_h(x, y) = l(y, h(x))$ for all $h \in \mathcal{H}$, and $l_{\mathcal{H}} = \{l_h : h \in \mathcal{H}\}$. In this section we will show how to obtain sample complexity bounds on algorithms for this basic learning problem using Theorem 6 above.

To obtain these bounds, we will need bounds on $\mathbf{dim_P}(l_{\mathcal{H}})$. In this section we will look at some useful tricks for computing $\mathbf{dim_P}(l_{\mathcal{H}})$ in the important case $A \subset \Re$, i.e. when each decision is represented by a real number. In the following section we discuss more general decision spaces.

When $A \subset \Re$, the functions in $\mathcal{H}$ are themselves real-valued, so we can talk about the pseudo dimension of $\mathcal{H}$ itself, without reference to any particular loss function. What makes this useful is that in many important cases the pseudo dimension of $\mathcal{H}$ is the same as the pseudo dimension of $l_{\mathcal{H}}$. Thus we can factor out the effects of the loss function in deriving our sample size bounds, and concentrate on the pseudo dimension of the hypothesis space $\mathcal{H}$.

Definition 5 *Let* $l : Y \times A \to \Re$ *be a loss function, where* $A \subset \Re$ *and* Y *is an arbitrary set. For each* $y \in Y$ *define the function* $f_y : A \to \Re$ *by letting* $f_y(a) = l(y, a)$ *for each* $a \in A$, *i.e.* f_y *is the restriction of* l *obtained by fixing its first argument to* y. *We say that* l *is* monotone *over* A *if for every* $y \in Y$, *either* f_y *is strictly increasing on* A, *or* f_y *is strictly decreasing on* A. *Thus* f_y *may be increasing for some* $y \in Y$ *and decreasing for others.*

Lemma 5 *If* $A \subset \Re$ *and* l *is a loss function on* $A \times Y$ *that is monotone over* A, *then* $\mathbf{dim_P}(l_{\mathcal{H}}) = \mathbf{dim_P}(\mathcal{H})$.

for all probability measures P on S and $m \geq d$. It is easily verified that if

$$\frac{M}{\epsilon} \ln(2\mathcal{M}(\epsilon, \mathbf{F}, d_{L^1(P)})) < d$$

then the upper bound given in Theorem 5 follows trivially from the fact that $\epsilon \leq M$. Thus we may assume that $\frac{M}{\epsilon} \ln(2\mathcal{M}(\epsilon, \mathbf{F}, d_{L^1(P)})) \geq d$. Hence, if $m \geq \frac{M}{\epsilon} \ln(2\mathcal{M}(\epsilon, \mathbf{F}, d_{L^1(P)}))$, then $m \geq d$ and

$$(1 - \mathcal{M}(\epsilon, \mathbf{F}, d_{L^1(P)}) e^{-\epsilon m/M}) \geq 1/2.$$

Thus from (4) we obtain

$$\left(\frac{eM \ln(2\mathcal{M}(\epsilon, \mathbf{F}, d_{L^1(P)}))}{\epsilon d} \right)^d \geq \frac{1}{2} \mathcal{M}(\epsilon, \mathbf{F}, d_{L^1(P)}).$$

With some simple calculations, this gives the bound of Theorem 5. □

Using our results on uniform convergence from sections 3.2 and 10.4, we can now show the following.

Theorem 6 *Let* $\mathbf{F}$ *be a permissible family of functions from a set* S *into* $[0, M]$ *with* $\mathbf{dim_P}(\mathbf{F}) = d$ *for some* $1 \leq d < \infty$. *Assume* $m \geq 1$, $0 < \nu \leq 8M$ *and* $0 < \alpha < 1$. *Let* $\bar{\xi}$ *be generated by* m *independent draws according to any distribution on* S. *Then*

$$\mathbf{Pr}\left(\exists f \in \mathbf{F} : d_\nu(\widehat{\mathbf{E}}_{\bar{\xi}}(f), \mathbf{E}(f)) > \alpha \right) \leq 8 \left(\frac{16eM}{\alpha\nu} \ln \frac{16eM}{\alpha\nu} \right)^d e^{-\alpha^2 \nu m / 8M}.$$

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. Let $\epsilon = \alpha\nu/8$. Since $\alpha < 1$ and $\nu \leq 8M$, $\epsilon \leq M$. 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 measure induced by $\bar{\xi}$, in which each set has measure equal to the fraction of the points in $\bar{\xi}$ it contains. Thus by Theorem 11 of section 10.1 and Theorem 5, we have

$$\mathcal{N}(\epsilon, \mathbf{F}_{|\bar{\xi}}, d_{L^1}) \leq \mathcal{M}(\epsilon, \mathbf{F}_{|\bar{\xi}}, d_{L^1}) \leq 2 \left(\frac{2eM}{\epsilon} \ln \frac{2eM}{\epsilon} \right)^d,$$

for all $\bar{\xi} \in S^{2m}$. Hence the given probability is at most

$$8 \left(\frac{2eM}{\epsilon} \ln \frac{2eM}{\epsilon} \right)^d e^{-\alpha^2 \nu m / 8M} = 8 \left(\frac{16eM}{\alpha\nu} \ln \frac{16eM}{\alpha\nu} \right)^d e^{-\alpha^2 \nu m / 8M}$$

by Theorem 12.

For the second result, setting the bound above equal to δ and solving for m gives

$$m \geq \frac{8M}{\alpha^2 \nu} \left(d \ln \left(\frac{16eM}{\alpha\nu} \ln \frac{16eM}{\alpha\nu} \right) + \ln \frac{8}{\delta} \right).$$

It is easily verified that $\ln(a \ln a) < 2 \ln(a/2)$ when $a \geq 5$, and from this the bound given in the second result follows. □

independently at random from the uniform distribution on $[0, M]$. *Let* $\bar{\xi} = (\xi_1, \ldots, \xi_m)$ *be a random vector in* S^m *where each* ξ_i *is drawn independently at random from* P. *Then for all* $\epsilon > 0$

$$\mathbf{E}(|sign(\mathbf{F}_{|\bar{\xi}} - \bar{r})|) \geq \mathcal{M}(\epsilon, \mathbf{F}, d_{L^1(P)}) \left(1 - \mathcal{M}(\epsilon, \mathbf{F}, d_{L^1(P)}) e^{-\epsilon m/M}\right).$$

Proof. For all $f \in \mathbf{F}$ we will denote $(f(\xi_1), \ldots, f(\xi_m))$ by $f_{|\bar{\xi}}$. Choose $\epsilon > 0$. Let $\mathbf{G}$ be an ϵ-separated subset of $\mathbf{F}$ (w.r.t. $d_{L^1(P)}$), with $|\mathbf{G}| = \mathcal{M}(\epsilon, \mathbf{F}, d_{L^1(P)})$. Then

$$\begin{aligned}
\mathbf{E}(|sign(\mathbf{F}_{|\bar{\xi}} - \bar{r})|) &\geq \mathbf{E}(|sign(\mathbf{G}_{|\bar{\xi}} - \bar{r})|) \\
&\geq \mathbf{E}(|\{f \in \mathbf{G} : sign(f_{|\bar{\xi}} - \bar{r}) \neq sign(g_{|\bar{\xi}} - \bar{r}) \text{ for all } g \in \mathbf{G},\ g \neq f\}|) \\
&= \sum_{f \in \mathbf{G}} \mathbf{Pr}\left(sign(f_{|\bar{\xi}} - \bar{r}) \neq sign(g_{|\bar{\xi}} - \bar{r}) \text{ for all } g \in \mathbf{G},\ g \neq f\right) \\
&= \sum_{f \in \mathbf{G}} \left(1 - \mathbf{Pr}\left(\exists g \in \mathbf{G},\ g \neq f : sign(f_{|\bar{\xi}} - \bar{r}) = sign(g_{|\bar{\xi}} - \bar{r})\right)\right) \\
&\geq \sum_{f \in \mathbf{G}} \left(1 - |\mathbf{G}|\, max_{g \in \mathbf{G}, g \neq f} \mathbf{Pr}\left(sign(f_{|\bar{\xi}} - \bar{r}) = sign(g_{|\bar{\xi}} - \bar{r})\right)\right).
\end{aligned}$$

Let f and g be distinct functions in $\mathbf{G}$. Since $\mathbf{G}$ is ϵ-separated,

$$\int_S |f(\xi) - g(\xi)| dP(\xi) > \epsilon.$$

In addition, the range of f and g is $[0, M]$. Hence if ξ_i is drawn at random from P and r_i drawn at random from the uniform distribution on $[0, M]$, then the probability that r_i lies between $f(\xi_i)$ and $g(\xi_i)$ is at least ϵ/M. And $sign(f_{|\bar{\xi}} - \bar{r}) = sign(g_{|\bar{\xi}} - \bar{r})$ only if this fails to occur for each i, $1 \leq i \leq m$. Thus

$$\mathbf{Pr}\left(sign(f_{|\bar{\xi}} - \bar{r}) = sign(g_{|\bar{\xi}} - \bar{r})\right) \leq \left(1 - \frac{\epsilon}{M}\right)^m \leq e^{-\epsilon m/M}.$$

Since this holds for every distinct f and g in $\mathbf{G}$, combining this with the inequality above, we have

$$\mathbf{E}(|sign(\mathbf{F}_{|\bar{\xi}} - \bar{r})|) \geq |\mathbf{G}|(1 - |\mathbf{G}| e^{-\epsilon m/M}).$$

Since $|\mathbf{G}| = \mathcal{M}(\epsilon, \mathbf{F}, d_{L^1(P)})$, this gives the result. □

Proof of Theorem 5. Since $\mathbf{dimp}(\mathbf{F}) = d$, it follows from Sauer's lemma that $|sign(\mathbf{F}_{|\bar{\xi}} - \bar{r})| \leq (em/d)^d$ for all $m \geq d$, $\bar{\xi} \in S^m$ and $\bar{r} \in [0, M]^m$. Hence the above lemma implies that

$$(em/d)^d \geq \mathcal{M}(\epsilon, \mathbf{F}, d_{L^1(P)}) \left(1 - \mathcal{M}(\epsilon, \mathbf{F}, d_{L^1(P)}) e^{-\epsilon m/M}\right) \tag{4}$$

The notions of ϵ-cover and metric dimension used in the previous section can be generalized to arbitrary pseudo metric spaces. This generalization is given in section 10.1 of the appendix. In the remainder of the paper we will use the concepts and notation given there without further special reference.

Using techniques that go back to Dudley [21], Pollard has obtained a beautiful theorem bounding the metric dimension of $(\mathbf{F}, d_{L^1(P)})$ by $\mathbf{dim_P(F)}$ for any probability measure P on S. Actually this result is much stronger in that it gives explicit bounds on the packing numbers for $\mathbf{F}$ using $d_{L^1(P)}$ balls of radius ϵ. Since packing numbers are closely related to covering numbers (Theorem 11 in section 10.1), these bounds can then be used with Theorem 2 to obtain uniform convergence results for empirical estimates of functions in $\mathbf{F}$. We now state and prove a version of Pollard's result ([60], Lemma 25, p. 27) for the special case when $\mathbf{F}$ is a class of functions taking values in the interval $[0, M]$ with somewhat better bounds on the packing numbers.

Theorem 5 *(Pollard) Let $\mathbf{F}$ be a family of functions from a set S into $[0, M]$, where $\mathbf{dim_P(F)} = d$ for some $1 \leq d < \infty$. Let P be a probability measure on S. Then for all $0 < \epsilon \leq M$,*

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

The proof we give uses essentially the same techniques as Pollard's, with some minor modifications. It relies on a few lemmas, which we give now. The first, which we give without proof, was discovered independently by a number of people (see [8]), including Vapnik and Chervonenkis [76], but is most often attributed to Sauer [67] in the computer science literature.

Lemma 3 *(Sauer) Let $\mathbf{F}$ be a class of functions from $S = \{1, 2, \ldots, m\}$ into $\{0, 1\}$ with $|\mathbf{F}| > 1$ and let d be the length of the longest sequence of points $\bar{\xi}$ from S such that $\mathbf{F}_{|\bar{\xi}} = \{0, 1\}^d$. Then* [12]

$$|\mathbf{F}| \leq \sum_{i=0}^{d} \binom{m}{i} \leq (em/d)^d$$

where e is the base of the natural logarithm. $\square$

In the next lemma we bound the packing numbers of $(\mathbf{F}, d_{L^1(P)})$ in terms of the expected number of orthants intersected by a random translation of a random restriction of $\mathbf{F}$. This is the key lemma of the proof.

Lemma 4 *Let $\mathbf{F}$ be a family of functions from a set S into $[0, M]$ and let P be a probability measure on S. Let $\bar{r} = (r_1, \ldots, r_m)$ be a random vector in $[0, M]^m$ where each r_i is drawn*

[12] See e.g. [22], Prop. 2.2.9, or [16], Appendix, for a proof of the second inequality. Note also that [76] actually contains a slightly weaker result.

The pseudo dimension has a few invariance properties that are useful (see [62] for further results of this type).

Theorem 4 *Let* $\mathbf{F}$ *be a family of functions from* S *into* $\Re$. *Fix any function* g *from* S *into* $\Re$ *and let* $\mathbf{G} = \{g + f : f \in \mathbf{F}\}$. *Let* I *be a real interval (possibly all of* $\Re$*) such that every function in* $\mathbf{F}$ *takes values only in* I. *Fix any nondecreasing (resp. nonincreasing) function* $h : I \to \Re$ *and let* $\mathbf{H} = \{h \circ f : f \in \mathbf{F}\}$, *where* $\circ$ *indicates function composition. Then*

1. *([82])* $\mathbf{dim_P}(\mathbf{G}) = \mathbf{dim_P}(\mathbf{F})$ *and*
2. *([56,23])* $\mathbf{dim_P}(\mathbf{H}) \leq \mathbf{dim_P}(\mathbf{F})$, *with equality if* h *is continuous and strictly increasing (resp. continuous and strictly decreasing).*

Proof. Part (1) follows directly from the fact that the notion of a set of points being full is invariant under translation. For part (2) it suffices to prove the results for h nondecreasing and h continuous and strictly increasing. Let $\bar{\xi} = (\xi_1, \ldots, \xi_d)$ be such that $\mathbf{H}_{|\bar{\xi}}$ is full, i.e. such that $\mathbf{H}_{|\bar{\xi}} - \bar{x}$ intersects all 2^d orthants of $\Re^d$ for some vector $\bar{x} = (x_1, \ldots, x_d)$ in $\Re^d$. Then for every Boolean vector $\bar{b} \in \{0,1\}^d$ there exists a function $f_{\bar{b}} \in \mathbf{F}$ such that for every i, $1 \leq i \leq d$, we have $h \circ f_{\bar{b}}(\xi_i) > x_i$ if and only if the i^{th} bit of $\bar{b}$ is 1. For each i, $1 \leq i \leq d$, let

$$u_i = min\{f_{\bar{b}}(\xi_i) : \textit{ the } i^{th} \textit{ bit of } \bar{b} \textit{ is } 1\}$$

and

$$l_i = max\{f_{\bar{b}}(\xi_i) : \textit{ the } i^{th} \textit{ bit of } \bar{b} \textit{ is } 0\}.$$

Since h is nondecreasing, we have $u_i > l_i$ for each i. Let $r_i = (u_i + l_i)/2$ for each i and $\bar{r} = (r_1, \ldots, r_d)$. Let $T = \{f_{\bar{b}} : \bar{b} \in \{0,1\}^d\}$. Then clearly $T - \bar{r}$ intersects every orthant of $\Re^d$, so T is full. Since $T \subset \mathbf{F}$, this implies that $\mathbf{F}_{|\bar{\xi}}$ is full, and hence $\mathbf{dim_P}(\mathbf{H}) \leq \mathbf{dim_P}(\mathbf{F})$. Equality follows when h is continuous and strictly increasing since we obtain the class $\mathbf{F}$ from $\mathbf{H}$ by composing with h^{-1}. □

By putting a probability measure on S, we can view a class $\mathbf{F}$ of real-valued functions on S as pseudo metric space. The distance between two functions is the integral of the absolute value of their difference, i.e. the L^1 distance, relative to the given measure. To make this work, we need to make some assumptions about the integrability of the functions in $\mathbf{F}$ under the given measure. Since we will be concerned only with families of functions taking values in a bounded range in this paper, this will cause no problems for us. For convenience, we choose this range to be $[0, M]$. For a more general treatment, see [60][22].

Definition 4 *Let* $\mathbf{F}$ *be a class of functions from* S *into* $[0, M]$, *where* $M > 0$, *and* P *be a probability measure on* S. *Then* $d_{L^1(P)}$ *is the pseudo metric on* $\mathbf{F}$ *defined by*

$$d_{L^1(P)}(f, g) = \mathbf{E}(|f - g|) = \int_S |f(\xi) - g(\xi)| dP(\xi) \textit{ for all } f, g \in \mathbf{F}.$$

to $\bar{y}$ and contains a strictly negative coordinate, giving an immediate contradiction, or $\bar{x}$ is non-zero and $\bar{x}$ is orthogonal to $\bar{y} - \bar{x}$. In this last case,

$$\sum_{i=1}^{d} x_i y_i = \sum_{i=1}^{d} x_i^2,$$

which is again a contradiction, since the left side is non-positive while the right side is strictly positive. □

It follows from this lemma that if T is contained in a hyperplane of $\Re^d$ then T is not full.

Definition 3 *Let* $\mathbf{F}$ *be a family of functions from a set* S *into* $\Re$. *For any sequence* $\bar{\xi} = (\xi_1, \ldots, \xi_d)$ *of points in* S, *let* $\mathbf{F}_{|\bar{\xi}} = \{(f(\xi_1), \ldots, f(\xi_d)) : f \in \mathbf{F}\}$. *If* $\mathbf{F}_{|\bar{\xi}}$ *is full then we say that* $\bar{\xi}$ *is* shattered *by* $\mathbf{F}$. *The* pseudo dimension *of* $\mathbf{F}$, *denoted* $\mathbf{dim_P(F)}$, *is the largest* d *such that there exists a sequence of* d *points in* S *that is shattered by* $\mathbf{F}$. *If arbitrarily long finite sequences are shattered, then* $\mathbf{dim_P(F)}$ *is infinite.*

It is clear that when $\mathbf{F}$ is a set of $\{0,1\}$-valued functions then for any sequence $\bar{\xi}$ of d points in S, $\mathbf{F}_{|\bar{\xi}}$ is full if and only if $\mathbf{F}_{|\bar{\xi}} = \{0,1\}^d$. Thus in this case $\mathbf{dim_P(F)}$ is the length d of the longest sequence of points $\bar{\xi}$ such that $\mathbf{F}_{|\bar{\xi}} = \{0,1\}^d$. This is the definition of the Vapnik-Chervonenkis dimension of a class $\mathbf{F}$ of $\{0,1\}$-valued functions [77][36][16]. Thus the pseudo dimension generalizes the Vapnik-Chervonenkis dimension to arbitrary classes of real-valued functions.

The pseudo dimension also generalizes the algebraic notion of the dimension of a vector space of real-valued functions [21].

Theorem 3 *(Dudley) Let* $\mathbf{F}$ *be a* d*-dimensional vector space of functions from a set* S *into* $\Re$. *Then* $\mathbf{dim_P(F)} = d$.

Proof. Fix any sequence $\bar{\xi} = (\xi_1, \ldots, \xi_{d+1})$ of points in S. For any $f \in \mathbf{F}$ let $\Psi(f) = (f(\xi_1), \ldots, f(\xi_{d+1}))$. Then Ψ is a linear mapping from $\mathbf{F}$ into $\Re^{d+1}$, and the image of Ψ is $\mathbf{F}_{|\bar{\xi}}$. Since $\mathbf{F}$ is a vector space of dimension d, this implies that $\mathbf{F}_{|\bar{\xi}}$ is a subspace of $\Re^{d+1}$ of dimension at most d. Hence by Lemma 2, $\mathbf{F}_{|\bar{\xi}}$ is not full. This implies $\mathbf{dim_P(F)} \leq d$. On the other hand, if $\mathbf{F}$ is a d-dimensional vector space of real-valued functions on S, then there exists a sequence $\bar{\xi}$ of d points in S such that $\mathbf{F}_{|\bar{\xi}} = \Re^d$. Hence $\bar{\xi}$ is shattered, implying that $\mathbf{dim_P(F)} \geq d$. □

There are many other ways that the VC dimension can be generalized to real-valued functions [53] [52] [60] [78] [23]. Dudley [23] compares several such generalizations, albeit in a different context. The generalization we have proposed here, the pseudo dimension, is a minor variant of the notion used by Pollard in [60] to define classes of real-valued functions of polynomial discrimination, called VC-subgraph classes in [23]. The pseudo dimension will be used in the form defined above in Pollard's new book [62].

The proof of Theorem 2 follows the proof of Pollard's Theorem 24 ([60], p. 25) in general outline. However, the use of the d_ν metric necessitates a number of substantial modifications. The approach taken here is different from that taken (independently, but prior to this work) by Pollard in [61]. Still different, and more involved, techniques are used in the more general theory of weighted empirical processes developed by Alexander [2,3]. While the proof of Theorem 2 we give is simpler, it is still somewhat lengthy, so it is given in Appendix 10.4.

4 Pseudo dimension of classes of real-valued functions

In this section we will look at one way that bounds on the covering numbers appearing in Theorem 2 can be obtained. This technique, due to Pollard [60], who extended methods from [21], is based on certain intuitions from combinatorial geometry. It generalizes the techniques based on the Vapnik-Chervonenkis dimension used in [16], which apply only to $\{0,1\}$-valued functions. We begin by establishing some basic notation.

Definition 2 *For $x \in \Re$, let $sign(x) = 1$ if $x > 0$ else $sign(x) = 0$. For $\bar{x} = (x_1, \ldots, x_d) \in \Re^d$, let $sign(\bar{x}) = (sign(x_1), \ldots, sign(x_d))$ and for $T \subset \Re^d$ let $sign(T) = \{sign(\bar{x}) : \bar{x} \in T\}$. For any Boolean vector $\bar{b} = (b_1, \ldots, b_d)$, $\{\bar{x} \in \Re^d : sign(\bar{x}) = \bar{b}\}$ is called the $\bar{b}$-*orthant *of $\Re^d$, where we have, somewhat arbitrarily, included points with value zero for a particular coordinate in the associated lower orthant. Thus $sign(T)$ denotes the set of orthants intersected by T. For any $T \subset \Re^d$, and $\bar{x} \in \Re^d$, let $T + \bar{x} = \{\bar{y} + \bar{x} : \bar{y} \in T\}$, i.e. the translation of T obtained by adding the vector $\bar{x}$. We say that T is* full *if there exists $\bar{x} \in \Re^d$ such that $sign(T + \bar{x}) = \{0,1\}^d$, i.e. if there exists some translation of T that intersects all 2^d orthants of $\Re^d$.*

The following result is well known and can be proved in a variety of ways. For example, it follows easily from well known bounds on the number of cells in arrangements of hyperplanes (see e.g. [25]). We give an elementary proof using a technique from [21].

Lemma 2 *No hyperplane in $\Re^d$ intersects all orthants of $\Re^d$.*

Proof. Let T be a hyperplane in $\Re^d$. Choose a vector $\bar{x} \in \Re^d$ as follows. If T includes the origin, then let $\bar{x}$ be any vector that is orthogonal to T and has at least one strictly negative coordinate. (For any nonzero orthogonal vector $\bar{x}$, if $\bar{x}$ doesn't have a negative coordinate then $-\bar{x}$ does.) Otherwise let $\bar{x}$ be the (nonzero) vector in T on the line perpendicular to T that passes through the origin. To complete the proof, we show that for all $\bar{y} \in T$, $sign(\bar{y}) \neq \bar{1} - sign(\bar{x})$, where $\bar{1}$ denotes the all 1's vector.

Suppose to the contrary that $sign(\bar{y}) = \bar{1} - sign(\bar{x})$ for some $\bar{y} \in T$. This implies that the inner product $\sum_{i=1}^{d} x_i y_i$ is non positive, and is in fact strictly negative if either $\bar{x}$ or $\bar{y}$ contain a strictly negative coordinate. However, by our choice of $\bar{x}$, either $\bar{x}$ is orthogonal

samples $\bar{\xi}$, and hence the random covering number $\mathbf{E}(\mathcal{N}(\epsilon, \mathbf{F}_{|\bar{\xi}})) \leq |\mathbf{F}|$ for all ϵ, all sample sizes m, and all distributions on S. The main result about uniform empirical estimates for infinite classes of functions is similar to Theorem 1 except that the random covering numbers are used in place of $|\mathbf{F}|$.

Theorem 2 *([61]) Let* $\mathbf{F}$ *be a permissible*[11] *set of functions on* S *with* $0 \leq f(\xi) \leq M$ *for all* $f \in \mathbf{F}$ *and* $\xi \in S$. *Let* $\bar{\xi} = (\xi_1, \ldots, \xi_m)$ *be a sequence of* m *examples drawn independently from* S *according to any distribution on* S. *Then for any* $\nu > 0$ *and* $0 < \alpha < 1$,

$$\mathbf{Pr}\left(\exists f \in \mathbf{F} : d_\nu(\hat{\mathbf{E}}_{\bar{\xi}}(f), \mathbf{E}(f)) > \alpha\right) \leq 4\mathbf{E}\left(\mathcal{N}(\alpha\nu/8, \mathbf{F}_{|\bar{\xi}})\right) e^{-\alpha^2\nu m/16M}. \ \square$$

Corollary 1 *([60]) Under the same assumptions as above, for all* $\epsilon > 0$,

$$\mathbf{Pr}\left(\exists f \in \mathbf{F} : |\hat{\mathbf{E}}_{\bar{\xi}}(f) - \mathbf{E}(f)| > \epsilon\right) \leq 4\mathbf{E}\left(\mathcal{N}(\epsilon/16, \mathbf{F}_{|\bar{\xi}})\right) e^{-\epsilon^2 m/128M^2}.$$

Proof of corollary. This follows directly from the above result by setting $\nu = 2M$, and $\alpha = \epsilon/4M$. To see this, note that property (3) of the d_ν metric (section 2.2) implies that $|r - s| \leq \epsilon$ whenever $d_\nu(r, s) \leq \alpha$ for all $0 \leq r, s \leq M$ when this setting of ν and α is used. $\square$

The constants in these results are only crude estimates. No serious attempt has been made to minimize them.

The bound in this latter result depends critically on the relative magnitudes of the negative exponent in $e^{-\epsilon^2 m/128M^2}$ and the exponent in the expectation of the covering number $\mathcal{N}(\epsilon/16, \mathbf{F}_{|\bar{\xi}})$, which reflects the extent to which $\mathbf{F}_{|\bar{\xi}}$ "fills up" the m-cube $[0, M]^m$. For example, if $\mathbf{F}_{|\bar{\xi}}$ has metric dimension at most n for all m and all $\bar{\xi}$, then there is a constant c_0 such that for any $\eta > 0$, $\mathcal{N}(\epsilon/16, \mathbf{F}_{|\bar{\xi}}) \leq (c_0 M/\epsilon)^{n+\eta}$ for suitably small ϵ. In this case the negative exponential term eventually dominates the expected covering number, and beyond a critical sample size

$$m_0 = O\left(\frac{nM^2}{\epsilon^2} \log \frac{M}{\epsilon}\right),$$

the bound goes to zero exponentially fast. We will see examples of this in the following section, where we give bounds on the metric dimension of $\mathbf{F}_{|\bar{\xi}}$ in terms of a combinatorial parameter called the pseudo dimension of $\mathbf{F}$. The theorem actually shows that this exponential drop off occurs even if this metric dimension bound holds only for "most" $\bar{\xi}$.

On the other hand, if with high probability $\mathbf{F}_{|\bar{\xi}}$ "fills up" the m-cube $[0, M]^m$ to the extent that $\mathcal{N}(\epsilon/16, \mathbf{F}_{|\bar{\xi}}) \approx (c_0/\epsilon)^m$, which is as large as possible, then the covering number dominates, and the bound is trivial. Results in [77] (Theorem A.2, page 220) indicate that uniform convergence does not take place in this case. Similar remarks apply to the bound given in Theorem 2, which uses the d_ν metric.

[11] This is a measurability condition defined in section 10.2 of the appendix.

with a finite class $\mathbf{F}_0$ that "approximates" it, in the sense that each function in $\mathbf{F}$ is close to some function in $\mathbf{F}_0$, and argue that some type of uniform convergence of empirical estimates for $\mathbf{F}_0$ implies uniform convergence for $\mathbf{F}$. In the simplest version of this technique, the choice of $\mathbf{F}_0$ depends only on $\mathbf{F}$ and the distribution P, as in the "direct method" discussed in section II.2 of [60] (see also [77] section 6.6, [22] chapter 6, [13], [83]). However, more general results (apart from certain measurability constraints) are obtained by allowing $\mathbf{F}_0$ to depend on the particular random sample $\bar{\xi}$ (e.g. [60], chapter 2). Here $\mathbf{F}_0$ is called a "random cover", and its size is called a "random covering number". It is this type of result that we derive here.

We will need a few preliminary definitions to introduce the notion of ϵ-covers and metric dimension. A more general treatment of these ideas is given in the appendix, section 10.1. This more general treatment will be used later, but the following definitions suffice for this section.

For any real vectors $\bar{x} = (x_1, \ldots, x_m)$ and $\bar{y} = (y_1, \ldots, y_m)$ in $\Re^m$, let $d_{L^1}(\bar{x}, \bar{y}) = \frac{1}{m}\sum_{i=1}^{m} |x_i - y_i|$. Thus d_{L^1} is the L^1 distance metric. Let T be a set of points that lie in a bounded region of $\Re^m$. For any $\epsilon > 0$, an *ϵ-cover* for T is a finite set $N \subset \Re^m$ (not necessarily contained in T) such that for all $\bar{x} \in T$ there is a $\bar{y} \in N$ with $d_{L^1}(\bar{x}, \bar{y}) \leq \epsilon$. The function $\mathcal{N}(\epsilon, T)$ denotes the size of the smallest ϵ-cover for T. We refer to $\mathcal{N}(\epsilon, T)$ as a *covering number*.

Following [40] we define the *upper metric dimension* of the set T of points by

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

The *lower metric dimension*, denoted by $\underline{\mathbf{dim}}$, is defined similarly using **liminf**. When $\overline{\mathbf{dim}}(T) = \underline{\mathbf{dim}}(T)$, then this quantity is denoted $\mathbf{dim}(T)$, and referred to simply as the *metric dimension* of T. Note that if $\mathcal{N}(\epsilon, T) = (g(\epsilon)/\epsilon)^n$, where $g(\epsilon)$ is polylogarithmic in $1/\epsilon$, then $\mathbf{dim}(T) = n$. Hence the metric dimension essentially picks out the exponent in the rate of growth of the covering number as a function of $1/\epsilon$.

Assume all functions in $\mathbf{F}$ map from S into $[0, M]$. For any sample $\bar{\xi} = (\xi_1, \ldots, \xi_m)$, with $\xi_i \in S$, let

$$\mathbf{F}_{|\bar{\xi}} = \{(f(\xi_1), \ldots, f(\xi_m)) : f \in \mathbf{F}\}.$$

We will call $\mathbf{F}_{|\bar{\xi}}$ the *restriction of* $\mathbf{F}$ *to* $\bar{\xi}$. Note that $\mathbf{F}_{|\bar{\xi}}$ is a set of points in the m-cube $[0, M]^m$. We can consider the size of the covering number $\mathcal{N}(\epsilon, \mathbf{F}_{|\bar{\xi}})$ as giving some indication of the "richness at scale $\approx \epsilon$" of the class $\mathbf{F}$ of functions, restricted to the domain $\xi_1, \ldots, \xi_m$. The metric dimension of $\mathbf{F}_{|\bar{\xi}}$ gives some indication the "number of essential degrees of freedom" in this restriction of $\mathbf{F}$.

When $\xi_1, \ldots, \xi_m$ are drawn independently at random from S, the *random covering number* $\mathbf{E}(\mathcal{N}(\epsilon, \mathbf{F}_{|\bar{\xi}}))$ gives some indication of the "richness" of $\mathbf{F}$ on a "typical" set of m points in the domain S. Note that for finite $\mathbf{F}$, we have $\mathcal{N}(\epsilon, \mathbf{F}_{|\bar{\xi}}) \leq |\mathbf{F}|$ for all ϵ and all

Proof. For the second part of the theorem, using Bernstein's inequality (see e.g. [60]) it is easy to show that for any single function f with $0 \leq f \leq M$,

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

Details are given in Lemma 9, part (2) in the Appendix. It follows that the probability that there is any $f \in \mathbf{F}$ with $d_\nu(\hat{\mathbf{E}}_{\bar{\xi}}(f), \mathbf{E}(f)) > \alpha$ is at most $2|\mathbf{F}|e^{-\alpha^2 \nu m/M}$. Setting this bound to δ and solving for m gives the result on the sample size. The proof of the first part of the lemma is similar, except we use Hoeffding's inequality (see e.g. [60]), which implies that for any single f,

$$\mathbf{Pr}\left(|\hat{\mathbf{E}}_{\bar{\xi}}(f) - \mathbf{E}(f)| > \epsilon\right) \leq 2e^{-2\epsilon^2 m/M^2}.$$

□

By letting $\mathbf{F} = l_{\mathcal{H}}$, this theorem can be used in conjunction with Lemma 1 from the previous section to obtain bounds on the sample complexity of learning algorithms that minimize empirical error. Here we can use either the L_ϵ or $L_{\alpha,\nu}$ family of regret functions. In the former case we get a sample complexity

$$m(\epsilon, \delta) = O\left(\frac{M^2}{\epsilon^2}\left(\log|l_{\mathcal{H}}| + \log\frac{1}{\delta}\right)\right). \tag{1}$$

In the latter case we get a sample complexity

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

As shown in the previous section, a generalization of the PAC learning model can be obtained by using either the L_ϵ or $L_{\alpha,\nu}$ regret functions, in the latter case by setting $\alpha = 1/2$ and $\nu = \epsilon$. Note that plugging this latter setting into (2) gives a sample complexity

$$m(\epsilon, \delta) = O\left(\frac{M}{\epsilon}\left(\log|l_{\mathcal{H}}| + \log\frac{1}{\delta}\right)\right), \tag{3}$$

a significant improvement over (1), which is quadratic in M/ϵ. Thus the generalization of the PAC model using the d_ν metric to measure distance from optimality, and the resulting $L_{\alpha,\nu}$ family of regret functions, offers new insight in this regard. (Vapnik's use of the relative difference between empirical estimates and true expectations [77] also has this advantage; see [7] and the appendix of [16].)

3.2 The general case

The main task of this section is to generalize Theorem 1 to infinite collections of uniformly bounded functions. The basic idea is simple: we replace the infinite class $\mathbf{F}$ of functions

needed so that

$$\mathbf{Pr}\left(\exists f \in \mathbf{F} : |\widehat{\mathbf{E}}_{\bar{\xi}}(f) - \mathbf{E}(f)| > \epsilon\right) < \delta$$

for $\epsilon, \delta > 0$ [60] [22] [77]. Vapnik also obtains better bounds in some important cases by considering the relative deviation of empirical estimates from true expectations, i.e. he looks at bounds on the sample size needed so that

$$\mathbf{Pr}\left(\exists f \in \mathbf{F} : \frac{|\widehat{\mathbf{E}}_{\bar{\xi}}(f) - \mathbf{E}(f)|}{\mathbf{E}(f)} > \epsilon\right) < \delta.$$

As we mentioned in the previous section, this measure of deviation suffers from a discontinuity at $\mathbf{E}(f) = 0$, and lack of useful metric properties. As in [61], we will give bounds on the sample size needed so that

$$\mathbf{Pr}\left(\exists f \in \mathbf{F} : d_\nu(\widehat{\mathbf{E}}_{\bar{\xi}}(f), \mathbf{E}(f)) > \alpha\right) = \mathbf{Pr}\left(\exists f \in \mathbf{F} : \frac{|\widehat{\mathbf{E}}_{\bar{\xi}}(f) - \mathbf{E}(f)|}{\nu + \widehat{\mathbf{E}}_{\bar{\xi}}(f) + \mathbf{E}(f)} > \alpha\right) < \delta,$$

i.e. the deviation measured using the d_ν metric. By setting ν and α appropriately, we obtain results similar to those of [60] and [77] as special cases of our main theorem. However, our results are restricted to the case that all functions in $\mathbf{F}$ are positive and uniformly bounded.

3.1 The case of finite F

Before considering the general case, it is useful to see what bounds we can get in the case that $\mathbf{F}$ is a finite set of functions. Here we can easily prove the following.

Theorem 1 *Let $\mathbf{F}$ be a finite set of functions on S with $0 \leq f(\xi) \leq M$ for all $f \in \mathbf{F}$ and $\xi \in S$. Let $\bar{\xi} = (\xi_1, \ldots, \xi_m)$ be a sequence of m examples drawn independently from S according to any distribution on S, and let $\epsilon > 0$. Then*

$$\mathbf{Pr}\left(\exists f \in \mathbf{F} : |\widehat{\mathbf{E}}_{\bar{\xi}}(f) - \mathbf{E}(f)| > \epsilon\right) \leq 2|\mathbf{F}|e^{-2\epsilon^2 m/M^2}.$$

For $0 < \delta \leq 1$ and sample size

$$m \geq \frac{M^2}{2\epsilon^2}\left(\ln|\mathbf{F}| + \ln\frac{2}{\delta}\right)$$

this probability is at most δ. Further, for any $\nu > 0$ and $0 < \alpha < 1$,

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

For $0 < \delta \leq 1$ and sample size

$$m \geq \frac{M}{\alpha^2\nu}\left(\ln|\mathbf{F}| + \ln\frac{2}{\delta}\right)$$

this probability is at most δ. □

3 Uniformly good empirical estimates of means

In this section we concentrate on the problem of bounding the number of random examples needed to get good empirical estimates of the error of each of the hypotheses in a hypothesis space $\mathcal{H}$. For each hypothesis $h \in \mathcal{H}$ and example $\xi = (x, y) \in S$, let $l_h(\xi) = l(y, h(x))$. As in the previous section, we assume that l is a non-negative bounded loss function taking values in the interval $[0, M]$, thus for each hypothesis h, l_h defines a random variable taking values in $[0, M]$. The value of l_h on an example (x, y) is the loss incured when you use h to determine the action to take for instance x, and the outcome is y. The error of h is just the expectation of l_h, i.e.

$$\mathbf{er}_P(h) = \mathbf{E}(l_h) = \int_S l_h(\xi) dP(\xi).$$

Furthermore, if $\bar{\xi} = (\xi_1, \ldots, \xi_m)$ is a sequence of examples from S, then the empirical error of h on $\bar{\xi}$ is the empirical estimate of the mean of l_h based on the sample $\bar{\xi}$, which we denote by $\hat{\mathbf{E}}_{\bar{\xi}}(l_h)$, i.e.

$$\widehat{\mathbf{er}}_{\bar{\xi}}(h) = \hat{\mathbf{E}}_{\bar{\xi}}(l_h) = \frac{1}{m} \sum_{i=1}^{m} l_h(\xi_i).$$

Let $l_{\mathcal{H}} = \{l_h : h \in \mathcal{H}\}$. We need to draw enough random examples to get a uniformly good empirical estimate of the expectation of every random variable in $l_{\mathcal{H}}$.

The general problem of obtaining a uniformly good estimate of the expectation of every function in a class $\mathbf{F}$ of real-valued functions has been widely studied (see e.g. [77,60,22] and their references). If no assumptions at all are made about the functions in $\mathbf{F}$, we immediately run into the problem that some functions in $\mathbf{F}$ could take on arbitrarily large values with arbitrarily small probabilities, making it impossible to obtain uniformly good empirical estimates of all expectations with any finite sample size. This problem can be avoided by making assumptions about the moments of the functions in $\mathbf{F}$, as in [77], or by assuming that there exists a single non-negative function with a finite expectation (called an *envelope*) that lies above the absolute value of every function in $\mathbf{F}$, as in [60, 22]. In our case, when the loss takes only values in $[0, M]$, then the constant function M serves as an envelope. This case is especially nice since this same envelope works for all distributions on the domain S of the functions in $\mathbf{F}$.

The usual measure of deviation of empirical estimates from true means is simply the absolute value of the difference. Thus we would say that the empirical estimates for the expectations of the functions in $\mathbf{F}$ converge uniformly to the true expectations if as the size of the random sample $\bar{\xi}$ grows,

$$\mathbf{Pr}\left(\exists f \in \mathbf{F} : |\hat{\mathbf{E}}_{\bar{\xi}}(f) - \mathbf{E}(f)| > \epsilon\right)$$

goes to zero for any $\epsilon > 0$. (This is called *(uniform) convergence in probability*, see e.g. [15]). Vapnik, Dudley, Pollard and others have obtained general bounds on the sample size

at least $1 - \delta$, we have $\rho(\widehat{\mathbf{er}}_{\bar{\xi}}(h), \mathbf{er}_P(h)) \leq \epsilon$ for all $h \in \mathcal{H}$. Here ρ is some metric on $\Re^+$, e.g. either the absolute difference or the d_ν metric.

The following result shows that uniform convergence of the empirical error estimates, along with a learning function $\mathcal{A}$ that gives a randomized solution to the optimization problem on the estimates, gives a solution to the basic learning problem. We state it for the d_ν metric, but the same argument works also for the absolute difference metric.

Lemma 1 *Let $\nu > 0$ and $0 < \alpha, \delta < 1$. Suppose the sample size $m = m(\alpha, \nu, \delta)$ is such that for all probability distributions $P \in \mathcal{P}$*

$$\mathbf{Pr}(\exists h \in \mathcal{H} : d_\nu(\widehat{\mathbf{er}}_{\bar{\xi}}(h), \mathbf{er}_P(h)) > \alpha/3) \leq \delta/2,$$

where the $\xi_i \in \bar{\xi}$, $1 \leq i \leq m$, are drawn independently at random from S according to the distribution P. Suppose also that the algorithm $\mathcal{A}$ is such that for all $P \in \mathcal{P}$

$$\mathbf{Pr}(d_\nu(\widehat{\mathbf{er}}_{\bar{\xi}}(\mathcal{A}(\bar{\xi})), \widehat{\mathbf{er}}^*_{\bar{\xi}}) > \alpha/3) \leq \delta/2,$$

where $\bar{\xi}$ is drawn randomly by P as above. Then for all $P \in \mathcal{P}$

$$\mathbf{Pr}(d_\nu(\mathbf{er}_P(\mathcal{A}(\bar{\xi})), \mathbf{er}^*_P) > \alpha) \leq \delta,$$

i.e. $\mathcal{A}$ solves the basic learning problem for the family of $L_{\alpha,\nu}$ regret functions and has sample complexity at most $m(\alpha, \nu, \delta)$.

Proof. By the triangle inequality for d_ν, if

1. $d_\nu(\mathbf{er}_P(\mathcal{A}(\bar{\xi})), \widehat{\mathbf{er}}_{\bar{\xi}}(\mathcal{A}(\bar{\xi}))) \leq \alpha/3$,
2. $d_\nu(\widehat{\mathbf{er}}_{\bar{\xi}}(\mathcal{A}(\bar{\xi})), \widehat{\mathbf{er}}^*_{\bar{\xi}}) \leq \alpha/3$, and
3. $d_\nu(\widehat{\mathbf{er}}^*_{\bar{\xi}}, \mathbf{er}^*_P) \leq \alpha/3$,

then

$$d_\nu(\mathbf{er}_P(\mathcal{A}(\bar{\xi})), \mathbf{er}^*_P) \leq \alpha.$$

The second assumption of the lemma states that (2) holds with probability at least $1-\delta/2$. The first assumption implies that both (1) and (3) hold with probability at least $1 - \delta/2$. (If (3) fails then we can find a hypothesis $h \in \mathcal{H}$ such that $d_\nu(\widehat{\mathbf{er}}_{\bar{\xi}}(h), \mathbf{er}_P(h)) > \alpha/3$. Here we use the compatibility of d_ν with the ordering on the reals.) Hence with probability at least $1 - \delta$ all of (1) - (3) hold. The result follows. □

In statistics, this type of result is called a *consistency theorem* about the "statistic" (i.e. the hypothesis) computed by the learning function $\mathcal{A}$. This use of the term "consistency" differs sharply from that common in PAC learning research.

$\hat{h}$ produced by the algorithm $\mathcal{A}$ satisfies

$$\mathbf{er}_{P,l}(\hat{h}) \leq \mathbf{er}^*_{P,l} + \epsilon,$$

i.e. the error of $\hat{h}$ is at most ϵ greater than that of the optimal hypothesis in $\mathcal{H}$. When $L = L_{\alpha,\nu}$, this condition is the same, except that we require

$$\mathbf{er}_{P,l}(\hat{h}) \leq \frac{1+\alpha}{1-\alpha}\mathbf{er}^*_{P,l} + \frac{\alpha\nu}{1\ -\ \alpha}.$$

Thus in the former case, the sample complexity is defined in terms of small additive deviations from optimality, and in the latter, we allow both additive and multiplicative deviations. These deviations are controlled by the parameters α and ν.

For example, when $\mathbf{er}^*_{P,l} = 0$ as in the standard PAC model, then setting $\alpha = 1/2$ and $\nu = \epsilon$ makes the L_ϵ and $L_{\alpha,\nu}$ conditions equivalent; each reduces to the PAC condition

$$\mathbf{er}_{P,l}(\hat{h}) \leq \epsilon.$$

When $\mathbf{er}^*_{P,l} > 0$, then $L_{\alpha,\nu}$ condition approximates the L_ϵ condition when α is small and $\nu \approx \epsilon/\alpha$. In particular, since we are assuming that the underlying loss function l is bounded between 0 and M, we have $0 \leq \mathbf{er}_{P,l}(\hat{h}), \mathbf{er}^*_{P,l} \leq M$ and property (3) of the d_ν metric shows that the $L_{\alpha,\nu}$ condition with $\nu = 2M$ and $\alpha = \epsilon/4M$ implies the L_ϵ condition. Hence, the $L_{\alpha,\nu}$ condition is generally more flexible.

2.5 Relation between learning and optimization

Let us assume that the underlying loss function l is fixed, and we are using either the L_ϵ or $L_{\alpha,\nu}$ regret functions derived from l. In order to solve a basic learning problem, we must find, with high probability, a hypothesis $\hat{h}$ with error close to optimal. As the true distribution P is unknown, to do this we must rely on estimates of $\mathbf{er}_P(h)$ for the various $h \in \mathcal{H}$ which are derived from the given random training sample. For a given $h \in \mathcal{H}$ and training sample $\bar{\xi} = (\xi_1, \ldots, \xi_m)$, where $\xi_i = (x_i, y_i) \in S$, let $\widehat{\mathbf{er}}_{\bar{\xi}}(h)$ denote the empirical error on $\bar{\xi}$, i.e. $\widehat{\mathbf{er}}_{\bar{\xi}}(h) = \frac{1}{m}\sum_{i=1}^m l(y_i, h(x_i))$. Let $\widehat{\mathbf{er}}^*_{\bar{\xi}} = \inf\{\widehat{\mathbf{er}}_{\bar{\xi}}(h) : h \in \mathcal{H}\}$. We can then define a natural optimization problem associated with the basic learning problem: given the training sample $\bar{\xi}$, find a hypothesis $\hat{h} \in \mathcal{H}$ such that $\widehat{\mathbf{er}}_{\bar{\xi}}(\hat{h})$ is close to $\widehat{\mathbf{er}}^*_{\bar{\xi}}$, i.e. a hypothesis whose empirical error on the training sample is close to minimal.

Solving the optimization problem does not automatically solve the learning problem. We need to have good empirical error estimates as well. Since l is bounded, for every $h \in \mathcal{H}$, as the sample size $m \to \infty$, $\widehat{\mathbf{er}}_{\bar{\xi}}(h) \to \mathbf{er}_P(h)$ with probability 1. We will say that the empirical error estimates of hypotheses in $\mathcal{H}$ converge *uniformly* to the true error rates if for all ϵ and $\delta > 0$ there exists a sample size m such that when the $\xi_i \in \bar{\xi}$, $1 \leq i \leq m$, are drawn independently at random from S according to the distribution P, with probability

we can define a notion of "average case" risk to be minimized. The former goal is know as *minimax optimality*, and has been used in the PAC model. The later is the *Bayesian* notion of optimality [14,39], and has been used in several approaches to learning in neural nets based on statistical mechanics [18,72,30,71,58]. Unfortunately this last question has no clear cut answer, and leads us directly into a longstanding unresolved debate in statistics (see e.g. [44] and following discussion.). Since we have set out to generalize the PAC model, and since our results are best illustrated in the minimax setting, we will formalize the notion of a basic learning problem using the minimax criterion. However, as with our choice of regret function, this choice should not be interpreted as a dogmatic statement that this is the only "correct" model.

After making this difficult choice, we can now define exactly what we mean by a basic learning problem, and what it means for a learning algorithm (or function) to solve this problem.

Definition 1 *A basic learning problem is defined by six components X, Y, A, $\mathcal{H}$, $\mathcal{P}$, and $\mathcal{L}$, where the first five components are as defined in section 2.1, and the last component, $\mathcal{L}$, is a family of regret functions as defined in section 2.3 (e.g. $\mathcal{L} = \{L_{\alpha,\nu} : \nu > 0 \text{ and } 0 < \alpha < 1\}$, or $\mathcal{L} = \{L_\epsilon : \epsilon > 0\}$ for some loss function l). Let $\mathcal{A}$ be a learning function as defined in section 2.3. We say that $\mathcal{A}$ solves the basic learning problem if for all $L \in \mathcal{L}$ and all $0 < \delta < 1$ there exists a finite sample size $m = m(L, \delta)$ such that for all $P \in \mathcal{P}$,*

$$R(L, P, \mathcal{A}, m) \leq \delta.$$

The sample complexity *of the learning function $\mathcal{A}$ is the smallest such integer valued function $m(L, \delta)$. When $\mathcal{L} = \{L_{\alpha,\nu} : \nu > 0 \text{ and } 0 < \alpha < 1\}$ we will denote $m(L_{\alpha,\nu}, \delta)$ by $m(\alpha, \nu, \delta)$ and when $\mathcal{L} = \{L_\epsilon : \epsilon > 0\}$ we will denote $m(L_\epsilon, \delta)$ by $m(\epsilon, \delta)$.*

As discussed above, this definition generalizes the PAC criterion, and several others as well. In fact, this definition is quite generous, in that sample size needed to get the risk less than δ is only required to be finite for each $\delta > 0$. In particular, using property (3) of the d_ν metric from section 2.2, when the underlying loss function l is bounded, as we assume here, any algorithm $\mathcal{A}$ solves the basic learning problem using the $L_{\alpha,\nu}$ class of regret functions if and only if it solves it using the L_ϵ class. Thus it doesn't matter which of these two classes of regret functions we use. However, in practice it is the sample complexity of $\mathcal{A}$ that is critical, and this will depend on which class of regret functions are used.

The nature of this dependence is seen more clearly when we expand the condition

$$R(L, P, \mathcal{A}, m) \leq \delta$$

for $L = L_\epsilon$ and $L = L_{\alpha,\nu}$. When $L = L_\epsilon$, this condition means that given m random training examples drawn according to P, with probability at least $1 - \delta$, the hypothesis

demanding risk at most δ gives the usual PAC criterion that the error of the hypothesis produced by $\mathcal{A}$ be greater than ϵ with probability at most δ.

The regret function L can also be defined similarly, but using the d_ν metric to measure distance from optimality, instead of the absolute difference. Specifically, for every $\nu > 0$ and $0 < \alpha < 1$, we can define the regret function $L_{\alpha,\nu}$ by letting $L_{\alpha,\nu}(P,h) = 1$ if $d_\nu(\mathbf{er}_{P,l}(h), \mathbf{er}^*_{P,l}) > \alpha$, and $L_{\alpha,\nu}(P,h) = 0$ otherwise. In this case the risk $R(L_{\alpha,\nu}, P, \mathcal{A}, m)$ measures the probability that the error of the hypothesis produced by the algorithm $\mathcal{A}$ has distance more than α from optimal in the d_ν metric, when the algorithm is given m random training examples drawn according to P. We will see in sections 2.4 and 3.1 why this sometimes gives a more useful and flexible definition of regret. In fact, in this paper, we will give our main results in terms of the family $\{L_{\alpha,\nu} : \nu > 0 \; and \; 0 < \alpha < 1\}$ of regret functions, and show how corresponding results may be derived as corollaries for the family $\{L_\epsilon : \epsilon > 0\}$ of regret functions.

Other regret functions are also possible and lead to different learning criteria. For example, another, perhaps simpler, way to define regret is to let $L(P,h) = \mathbf{er}_{P,l}(h) - \mathbf{er}^*_{P,l}$. When $\mathbf{er}^*_{P,l} = 0$, as it does in the standard noise-free PAC model, this definition makes the regret L equal to the error $\mathbf{er}_{P,l}$, i.e. the expectation of the underlying loss l. In this case the risk $R(L, P, \mathcal{A}, m)$ measures the expectation of the loss incurred by the learning algorithm $\mathcal{A}$ when it is given m random training examples drawn according to P, forms a hypothesis h, and then uses h to determine the action on one further independent random example drawn according to P. This gives a generalization of the learning criterion studied in [35]. When $\mathbf{er}^*_{P,l} \neq 0$, then the risk gives the expectation of the amount of such loss above and beyond the expected loss that would be suffered if the optimal hypothesis were used. In particular, in density estimation, where P and h are both densities on the instance space X, if $P \in \mathcal{H}$ then defining the regret by $L(P,h) = \mathbf{er}_{P,l}(h) - \mathbf{er}^*_{P,l}$ makes it equal to the Kullback-Leibler divergence from P to h. Hence the risk is the expected Kullback-Leibler divergence of the hypothesis h returned by the algorithm to the true density (see sections 1.1.3 and 1.1.4).

It is also possible to define the regret function L directly, without using an underlying loss function l. For example, in density estimation it is possible to use other measures of the distance between two densities, e.g. the Hellinger distance or the total variational distance, as in [10,85]. The criterion from [38] for inferring a good *model of probability* can also be defined using an appropriate regret function, without defining an underlying loss l.

2.4 Full formalization of the basic learning problem

Having defined the regret function, and thereby the risk function, we still face one last issue: do we want to minimize risk in the worst case over all possible states of nature P in $\mathcal{P}$, or do we want to assume a *prior distribution* on possible distributions in $\mathcal{P}$, so that

2.3 The regret function L and the risk R

Once we have specified how we measure closeness to optimality, we still need to specify our criteria for a successful learning algorithm. Do we need to have the error of the hypothesis found close to the optimum $\mathbf{er}^*_{P,l}$ with high probability, or should its average distance from $\mathbf{er}^*_{P,l}$ be small? Do we measure success in terms of the performance of the algorithm on the worst case distribution in $\mathcal{P}$, or do we use some average case analysis over distributions in $\mathcal{P}$? These questions lead us right back to decision theory again, but this time at a higher level in the analysis of learning.

To see this, consider the structure of a learning algorithm $\mathcal{A}$. For any sample size m, the algorithm $\mathcal{A}$ may be given a sample $\bar{\xi} = ((x_1, y_1), \ldots, (x_m, y_m))$ drawn at random from S^m according an unknown product distribution P^m, where $P \in \mathcal{P}$. For any such $\bar{\xi}$ it will choose a hypothesis $\mathcal{A}(\bar{\xi}) \in \mathcal{H}$. Thus abstractly, the algorithm defines a function $\mathcal{A}$ from the set of all samples over S into $\mathcal{H}$, i.e. $\mathcal{A} : \bigcup_{m \geq 1} S^m \to \mathcal{H}$. Since we are not requiring computability here, we will call such $\mathcal{A}$ a *learning function*. When $P \in \mathcal{P}$ is the actual "state of nature" governing the generation of examples, and the algorithm produces the hypothesis $h \in \mathcal{H}$, let us say that we suffer a nonnegative real-valued *regret* $L(P, h)$. Thus, formally $L : \mathcal{P} \times \mathcal{H} \to \Re^+$. In our treatment here, the regret function L will be derived from the loss function l, and will measure the extent to which we have failed to produce a near optimal hypothesis, assuming P is the true state of nature (i.e. the amount of "regret" we feel for not having produced the optimal hypothesis). Finally, for each possible state of nature P, the average regret suffered by the algorithm, over all possible training samples $\bar{\xi} \in S^m$, is the *risk* of that algorithm under P for sample size m. This risk is defined formally by

$$R(L, P, \mathcal{A}, m) = \int_{\bar{\xi} \in S^m} L(P, \mathcal{A}(\bar{\xi}))dP^m(\bar{\xi}).$$

The goal of learning is to minimize risk.

We illustrate these definitions with a few examples. First suppose we want to capture the notion of successful learning that is used in the PAC model. Then one possibility is to introduce an *accuracy parameter* $\epsilon > 0$ and define the regret function $L = L_\epsilon$ by letting $L_\epsilon(P, h) = 1$ if $\mathbf{er}_{P,l}(h) - \mathbf{er}^*_{P,l} > \epsilon$, and $L_\epsilon(P, h) = 0$ otherwise. This we suffer regret only when the hypothesis h produced by the learning algorithm has error that is more than ϵ from optimal, measured by the absolute difference metric. For this definition of regret, the risk $R(L_\epsilon, P, \mathcal{A}, m)$ measures the probability that the hypothesis produced by $\mathcal{A}$ has error more than ϵ from optimal, when $\mathcal{A}$ is given m random training examples drawn according to P. We then demand that this risk be small, i.e. smaller than some given *confidence parameter* $\delta > 0$.

In the PAC model it is commonly assumed that the examples given to the algorithm $\mathcal{A}$ are noise-free examples of some underlying target function $f \in \mathcal{H}$. In this case the error $\mathbf{er}^*_P$ of the optimal hypothesis in $\mathcal{H}$ is zero, and hence $L_\epsilon(P, h) = 1 \Leftrightarrow \mathbf{er}_{P,l}(h) > \epsilon$. Hence

e.g. most measurements naturally have bounded ranges, but they can be annoying (see [78] and [60,62] for alternative approaches for unbounded loss functions).

2.2 Measuring distance from optimality with the d_ν metric

For a given hypothesis $h \in \mathcal{H}$ and distribution P on the sample space S, the expected loss of h is the average value of $l(y, h(x))$, when the example (x, y) is drawn at random according to P. It is defined by

$$\mathbf{er}_{P,l}(h) = \mathbf{er}_P(h) = \mathbf{E}(l(y, h(x))) = \int_S l(y, h(x)) dP(x, y)$$

(the subscript l will be omitted when the loss function is clear from the context.) Since l is bounded this expectation is finite for every distribution P. The expected loss $\mathbf{er}_P(h)$ is called the *error* of h, as in computational learning theory.

In section 1.1 we stated the goal of learning quite informally: Given examples chosen independently at random from some unknown probability distribution $P \in \mathcal{P}$, find a hypothesis $\hat{h}$ in $\mathcal{H}$ that comes "close to" minimizing the error $\mathbf{er}_P(h)$ over all $h \in \mathcal{H}$. Let $\mathbf{er}^*_{P,l}$ (or $\mathbf{er}^*_P$ when l is clear from the context) denote the infimum of $\mathbf{er}_P(h)$ over all h in the hypothesis space $\mathcal{H}$. To formalize our notion of a basic learning problem, we first need to say what we mean that $\mathbf{er}_P(\hat{h})$ is "close to" $\mathbf{er}^*_P$.

Let $r = \mathbf{er}_P(\hat{h})$ and $s = \mathbf{er}^*_P$. One natural interpretation is to demand that $|r - s| \leq \epsilon$ for some small $\epsilon > 0$. However, we will see in section 3.1 that sometimes it is better to use a relative measure of distance. For any real $\nu > 0$, let d_ν be the function defined by

$$d_\nu(r, s) = \frac{|r - s|}{\nu + r + s}$$

for any non-negative reals r and s. It is straightforward but tedious to verify that d_ν is a metric on $\Re^+$. The d_ν metric is similar to the standard function

$$\frac{|r - s|}{s}$$

used to measure the difference between the quality r of a given solution and the quality s of an optimal solution in combinatorial optimization. However, our measure has been modified to be well-behaved when one or both of its arguments are zero, and to be symmetric in its arguments (so that it is a metric). Three other properties of d_ν are also useful.

1. For all non-negative reals r and s, $0 \leq d_\nu(r, s) < 1$.
2. For all non-negative $r \leq s \leq t$, $d_\nu(r, s) \leq d_\nu(r, t)$ and $d_\nu(s, t) \leq d_\nu(r, t)$.
3. For $0 \leq r, s \leq M$, $\frac{|r-s|}{\nu+2M} \leq d_\nu(r, s) \leq \frac{|r-s|}{\nu}$.

We will refer to the second property by saying that d_ν is *compatible with the ordering on the reals.*

previous section. Finally, we will show how an algorithm can solve the learning problem by solving a related optimization problem.

2.1 The basic components $X, Y, A, \mathcal{H}, \mathcal{P}$ and l

We first review and further formalize the six components of the basic learning problem introduced in the previous section: X, Y, A, $\mathcal{H}$, $\mathcal{P}$ and l. The first four components are the instance, outcome, decision and hypothesis spaces, respectively. The first three of these are arbitrary sets, and the fourth, $\mathcal{H}$ is a family of functions from X into A. These have been discussed extensively in the previous section.

The fifth component, $\mathcal{P}$, is a family of joint probability distributions on $X \times Y$. These represent the possible "states of nature" that might be governing the generation of examples. The set $S = X \times Y$ will be called the *sample space*. We assume that examples are drawn independently at random according to some probability distribution $P \in \mathcal{P}$ on the sample space S. A sequence of examples will be called a *sample*. In what follows[9] we will usually assume that $\mathcal{P}$ includes all probability distributions on S. Hence our results will be distribution independent.

The last component, the loss function l, is a mapping from $Y \times A$ into $\Re$. In this paper we will assume that l is bounded and nonnegative, i.e. $0 \leq l \leq M$ for some real M. When Y and A are finite it is always possible to enforce this condition by simply adding a constant to l, which doesn't change the learning problem in any essential way. When either Y or A is infinite, the learning problem sometimes needs to be restricted to meet this condition. For example, in regression[10] we might restrict the possible parameter vectors in A and/or the possible outcomes in Y such that for every $y \in Y$ and $a \in A$, $\widehat{P}(y;a) \geq b$ for some constant b. We can then take $M = -\log b$. In density estimation, the same thing can be accomplished by restricting the instance space X to a bounded subset of $\Re^n$ on which all densities in $\mathcal{H}$ have values uniformly greater than b and less than B for constants $0 < b < B$. We can then add $\log B$ to the loss function to make it positive. The same method works for for estimating distributions on discrete spaces: we restrict ourselves to a finite instance space X and demand that for all $x \in X$ and all probability distributions $h \in \mathcal{H}$, $h(x) \geq b > 0$ (see e.g. [1,85]). These restrictions are often reasonable in practice,

[9] It is, however, possible and in fact common to assume that $\mathcal{P}$ is a very specific class of probability distributions on S. For example, let $X = \Re^n$. Then if we are doing classification learning and Y is discrete we may assume that y is selected according to an arbitrary distribution on Y, and for each y, $P(x|y)$ is a multi-dimensional Gaussian distribution on X [20]. On the other hand, if we are doing linear regression, then Y is real-valued and we might assume that x is selected according to an arbitrary distribution on X, and y is a linear function of x with additive Gaussian noise. In PAC learning theory we have a discrete analog of the latter case. Here we usually have $X = \{0,1\}^n$, $Y = \{0,1\}$, and y a Boolean function of x of a particular type (e.g. defined by a small disjunctive normal form formula), possibly plus random noise.

[10] Note that to get bounded loss in linear regression, X must a bounded subset of $\Re^n$ as well, since we can't bound Y without bounding X. The coefficients of the functions in $\mathcal{H}$ must also be bounded.

denotes the conditional distribution on Y, given $X = x$. The marginal distribution in X is defined by[8] $P_{|X}(x) = \sum_{y \in Y} P(x, y)$. Here and elsewhere, we abbreviate $P((x, y))$ by $P(x, y)$.

Finally, we list some other notation that is used several places in the paper, indicating which section it is defined in.

X,Y,A,$\mathcal{H}$ and l	sections 1.1 and 2.1
$\mathcal{P}$	section 2.1
$\mathbf{er}_{P,l}$, $\mathbf{er}_P$ (true error rates)	section 2.2
$\mathbf{er}^*_{P,l}$, $\mathbf{er}^*_P$ (optimal error rates)	section 2.2
$\widehat{\mathbf{er}}_{\bar{\xi}}$ (empirical error)	section 2.2
$\widehat{\mathbf{er}}^*_{\bar{\xi}}$ (optimal empirical error)	section 2.2
d_ν	section 2.2
L, L_ϵ, $L_{\alpha,\nu}$ (regret functions)	section 2.3
R (risk)	section 2.3
$m(\epsilon, \delta)$, $m(\alpha, \nu, \delta)$ (sample complexity)	section 2.4
$\mathcal{N}$ (covering number)	sections 10.1 and 3.2
$\mathcal{M}$ (packing number)	section 10.1
dim (metric dimension)	section 10.1
$\mathbf{dim_P}$ (pseudo dimension)	section 4
$\mathcal{C}$ (capacity)	section 6
ρ_l	section 6
$l_{\mathcal{H}}$	section 3
$\mathbf{F}_{\|\bar{\xi}}$	section 3
$\widehat{\mathbf{E}}$ (empirical expectation)	section 3
d_{L^1} (L^1 distance, vectors)	section 3.2
$d_{L^1(P)}$ (L^1 distance, functions)	section 4
$d_{L^1(P,\rho)}$ (L^1 distance, functions)	section 6

2 Learning and optimization

We now further formalize the basic problem of learning, as introduced in section 1.1. We will introduce a formal notion of a learning algorithm, and a higher level loss function, which we will call a *regret function*, that measures how well the learning algorithm performs. The regret function will be defined in terms of the low level loss function l discussed in the

[8] When S is uncountable, the marginal and conditional distributions are defined so that

$$\int_S f(x, y) dP(x, y) = \int_X \left(\int_Y f(x, y) dP(y|x) \right) dP_{|X}(x)$$

for every bounded measurable function f.

to make the paper more readable. The appendix has several sections. Section 10.1 contains a brief overview of the theory of metric spaces, ϵ-covers and metric dimension. Notation from this section is used in several places in the paper. Section 10.2 deals with certain technical measurability requirements. Section 10.3 gives an analogue of Chernoff and Hoeffding bounds using the d_ν metric. Section 10.4 contains the proof of Theorem 2. Finally, Section 10.5 contains a result on feedforward neural networks of linear threshold functions that is similar to that given in [12], and provides a counterpart to Corollary 3 in section 7.

1.5 Notational conventions

We denote the real numbers by $\Re$ and the non-negative real numbers by $\Re^+$. By log and ln we denote the logarithm base 2 and the natural logarithm, respectively. We use $\mathbf{E}(\cdot)$ to denote the expectation of a random variable, and $\mathbf{Var}(\cdot)$ to denote the variance of a random variable. When the probability space is defined implicitly from the context, we use $\mathbf{Pr}(\cdot)$ to denote the probability of a set. However, usually the measure on the underlying probability space will be defined explicitly using the symbol P.

In this paper, P will usually denote a probability measure on some appropriate[7] σ-algebra over the set $S = X \times Y$, where X is the instance space and Y is the outcome space. We use P^m to denote the m-fold product measure on S^m. Functions on S and subsets of S mentioned in the paper will be assumed to be measurable without explicit reference. Alternately, we will also view X and Y as random variables on some other, unspecified, probability space, e.g. when they are viewed as real valued measurements. In this case P is viewed as a joint distribution on X and Y. In either case, the probability of a set $T \subset S$ is defined by

$$P(T) = \int_T dP(\xi)$$

(where $\xi = (x, y)$ with $x \in X$ and $y \in Y$) and the expectation of function f on S is denoted by

$$\mathbf{E}(f) = \int_S f(\xi) dP(\xi).$$

When S is countable we will, with some abuse of notation, also use P for the probability mass function, i.e. for $\xi \in S$, $P(\xi)$ denotes $P(\{\xi\})$. Hence $P(T) = \sum_{\xi \in T} P(\xi)$ and $\mathbf{E}(f) = \sum_{\xi \in S} f(\xi) P(\xi)$ in this case. When S is continuous, a density associated with P (if it exists) is denoted by p.

When S is countable we use $P(y|x)$ to denote the probability that $Y = y$ given that $X = x$ (viewing X and Y as random variables) and similarly for $P(x|y)$. Hence $P(\cdot|x)$

[7] If S is countable then we assume this σ-algebra contains all subsets of S, otherwise we assume that S is a complete, separable metric space (see section 10.1) and that this σ-algebra is the smallest σ-algebra that contains the open sets of S (i.e. the σ-algebra of Borel sets).

study of fractals in nature [47], especially in connection with chaos in dynamical systems [27][28]. In this paper we build further on the beautiful results of Vapnik and Chervonenkis [77], Dudley [21] and Pollard [60], which relate a type of generalized VC dimension for a hypothesis space to the number of balls of radius ϵ required to cover the space, with respect to certain metrics. The sizes of the smallest such covers determine the *metric dimension* of the space. Our treatment closely parallels the approach given in [62]. It is interesting to note that related results connecting ϵ-covers with the VC dimension have also been independently developed in [13] and in recent computational geometry work [81][6]. This work seems to lead to a potentially rich area of investigation that combines elements of combinatorics, topology and geometry, and probability and measure in a novel framework. We feel that this area is not only fascinating from a purely mathematical standpoint, but also potentially very useful in machine learning and other applied fields.

1.4 Organization of the paper

The remainder of the paper is organized as follows. The learning framework we have described above in section 1.1 is defined more formally in section 2. There we also look at the question of evaluating the performance of learning algorithms in terms of the number of training examples they use. This question is also formalized from a decision theory perspective. We then provide a lemma (Lemma 1) that can be used to evaluate the performance of learning algorithms that work by minimizing empirical loss. To use this lemma, we need bounds on the rate of uniform convergence of empirical loss estimates to true expected losses. These are given in section 3. The key bound is given in Theorem 2 in section 3, and in a more general version in Theorem 12 in section 10.4 of the appendix.

To use the bound from Theorem 2 we need bounds on the "random covering numbers" associated with the hypothesis space $\mathcal{H}$, the loss function l and the distribution P. These are related to the idea of an ϵ-cover described above. In section 4 we introduce Pollard's notion of the pseudo dimension as a means of bounding the random covering numbers. Applications of this method to several learning problems are described in section 5.

Further tools for bounding the random covering numbers are developed in section 6. Here we introduce the notion of the capacity of the hypothesis space $\mathcal{H}$ (for a particular loss function l), and the related notion of the metric dimension of $\mathcal{H}$. In section 7 we use these notions to obtain bounds on the performance (in terms of the number of training examples used) of learning algorithms that use multilayer feedforward neural networks, and work by minimizing empirical loss (Corollary 3). Finally, some further discussion of our results is given in the conclusion, section 8.

Many of the more technical proofs and definitions have been moved into the appendix

[6] Specifically, Lemma 7.13 of [21] is nearly equivalent to Lemma 4.1 of [81] (using the primal space instead of the dual). This result also gives a stronger version of Theorem 4, part(3) of [13]. We give a still stronger version of this result, essentially due to Pollard, in Theorem 5 below.

in practice, e.g. as explicit formulae for choosing an appropriate sample size. Cross validation techniques, in which some of the training examples are held in reserve and used instead to test the performance of the hypotheses produced by the learning algorithm, are likely to perform better for this task in practice (see e.g. [83,80]). Nevertheless, cross validation is only a means of estimating the amount of overfitting in the learning method in particular cases, i.e. it is only an engineering trick and provides no scientific explanation of the phenomenon. Our goal is to understand and explain overfitting in general hypothesis spaces, from a scientific rather than an engineering viewpoint.

Finally, we should note that in practice, many learning algorithms do more than just search for a hypothesis in a fixed hypothesis space that minimizes empirical loss. For example, it is common to let the hypothesis space depend on the number of training examples available, using richer and richer hypothesis spaces as more examples become available (see e.g. [83,16]). This can allow the learning algorithm to produce a sequence of hypotheses with expected losses that approach the loss of Bayes optimal hypothesis in the limit of infinite training sample size for a large class of possible joint distributions. The results in this paper can be used to estimate the appropriate rate at which the hypothesis space should grow relative to the sample size to avoid overfitting. Other approaches, e.g. the method of *structural risk minimization* introduced by Vapnik [77], the *minimum description length* (MDL) approach [10,65], and the work using weight penalty functions in neural net training [79,43], try to find a hypothesis that minimizes some function of empirical loss and hypothesis complexity. These can also achieve Bayes optimal performance in the limit, and may be more effective in practice. Although uniform convergence results such as those we develop here are also be used in the analysis of such methods [77] (and in the analysis of cross-validation methods [56]), the full treatment of such approaches is beyond the scope of the present paper.

1.3 Overview of methods used

We now briefly discuss the methodology and previous work used in obtaining our results. Our work builds directly on the work of Vapnik and Chervonenkis, Pollard, and Dudley on the uniform convergence of empirical estimates [77][60][22] and its application to pattern recognition [77,78] [19]. It also builds on the work of Benedek and Itai on PAC learnability with respect to specific probability distributions [13], and is related to the work of Natarajan and Tadepalli on extensions of the VC dimension to multi-valued functions [54] [53] and PAC learnability with respect to classes of probability distributions [51] [52]. In addition, Quiroz and Kulkarni have each independently generalized the PAC model in a related, but more restrictive manner [63,41].

One of the key ideas we use is the notion of an ϵ-cover of a metric space [22] [60] [13] [52] [63] and the associated idea of *metric dimension* [40] (also called the *fractal* dimension [27]). This notion of dimension has played an important role in the now very active

In order to focus on the problem of overfitting, we take a simplified view of learning, in which the learner chooses a hypothesis space $\mathcal{H}$, and then tries to find a hypothesis in $\mathcal{H}$ with near minimal expected loss. To do this, the learner looks for a hypothesis that minimizes the observed average loss on the training examples, which is called *empirical loss* or *empirical error*. For example, in standard linear regression[5] the learning algorithm is the method of least squares, i.e. we find the linear function h that minimizes the average of $l(y, h(x)) = (h(x) - y)^2$ over all examples (x, y) in our training set. It is well known that if we have too few training examples, then we tend to overfit them, and the function we find does not come close to minimizing the actual expected quadratic loss, which would be obtained by integrating over all possible (mostly unseen) examples with respect to the unknown joint distribution on them. This same situation occurs with all nontrivial hypothesis spaces.

Using certain measures of the "dimension" or "capacity" of the hypothesis space $\mathcal{H}$ and classes derived from $\mathcal{H}$ (see below), we obtain general results on the number of random training examples needed so that with high probability, any hypothesis in $\mathcal{H}$ that has small empirical loss on the training examples will have small actual expected loss, i.e. we get uniform convergence results for empirical estimates like those in [77][22][60,62]. We show how these give upper bounds on sufficient training sample size like those derived in [16] and elsewhere using the notion of the VC dimension, and generalize those results.

As an application, we give specific bounds on the number of training examples needed to avoid overfitting when learning with the hypothesis space of feedforward neural nets [66], extending previous work in [12] and [83] (see also related work in [7]). These are the nets most widely used in current neural net learning research. Our model for feedforward neural nets is quite general in that it allows many types of units in the nets, including quasi-linear units [66], radial basis units [59], and product units [24].

In our general setting, successful learning means finding a hypothesis with average loss close to minimal over all hypotheses in the given hypothesis space, rather than loss close to zero as in the PAC model. In addition to using an additive model as in [45], we also define "close to" using a measure of relative difference (the d_ν *metric*) similar to the standard multiplicative measure of approximation used in combinatorial optimization. This allows us to state the relevant uniform convergence bounds as generalized "Chernoff-style" [6] bounds, as in [61],[17] (chapter 12), rather than "Hoeffding-style" bounds (as in Pollard's results [60]), giving better bounds on sufficient training sample size in some important cases.

We give these upper bounds on required sample size only to give some indication of the order-of-magnitude dependence of sample size on certain critical parameters of the learning problem, and to illustrate the theory. They are still too crude to be used directly

[5] For general regression with the negative log likelihood loss function, the principle of minimizing empirical loss is the same as the principle of *maximum likelihood* [14,39].

pursue this here.

1.2 Summary and discussion of the results presented in this paper

There are three major practical issues in this decision theoretic view of learning. The first is the number of random examples needed in order to be able to produce a good hypothesis in the hypothesis space $\mathcal{H}$, i.e. a hypothesis whose expected loss is near the minimum of all hypotheses in $\mathcal{H}$. If too few examples are used, we run into the problem of *overfitting*, where the hypothesis produced performs well on the training data, but not on further random examples drawn from the same joint distribution that generated this training data. The second is the adequacy of the hypothesis space $\mathcal{H}$. If $\mathcal{H}$ does not contain any hypothesis with expected loss close to that of Bayes optimal hypothesis for the particular joint distribution we are dealing with, then we can never hope to achieve near optimal performance using this hypothesis space. Choosing the right hypothesis space often requires considerable insight into the particular problem domain. Finally, the third practical problem is the computational complexity of the method we use to produce our hypothesis from the training examples. This issue has been addressed extensively in the PAC literature, and is also addressed in [38,1]. Of these three important issues, we examine only the first in this paper. This issue is referred to as the problem of estimating the "sample complexity" of the learning problem in the PAC literature [26].

The number of random training examples needed to avoid overfitting depends critically on the nature of the hypothesis space used. Different kinds of hypothesis spaces are used in different areas of learning research, partly because different kinds of instance and outcome spaces are used. In pattern recognition and statistics, the instance space X is usually a finite dimensional real vector space, i.e. each instance consists of a vector of real valued measurements of some attributes. In density estimation, a hypothesis represents a density on X, and many choices are possible. One common choice is a mixture of Gaussian densities (e.g. [20][57]). In standard regression, the outcome and decision spaces Y and A are identical and real valued, and linear functions are most often used as hypotheses. For more complex outcome spaces such as those in the medical diagnosis example given above, the hypothesis space for regression is usually defined using a *generalized linear model* [48]. Similarly, in binary classification, where there are only two possible outcomes in Y as in the PAC model, linear threshold functions are most often used as hypotheses, and there are straightforward generalizations for the case of k-ary classification (see e.g. [20]). This "linear bias" in pattern recognition and statistics is in contrast to that in the PAC model and other AI areas, including work in neural networks, in which a rich variety of hypothesis spaces are used (see e.g. [73,74,31,32]). Our main goal in this paper is to develop analytic tools to help understand the problem of overfitting in these more complex hypothesis spaces.

hence this version of regression also fits naturally[4] into our proposed learning framework (see e.g. [84], [33]).

1.1.4 density and parameter estimation

Finally, the problems of parameter estimation and non-parametric density estimation can also be viewed as special cases of this decision theoretic framework. For parameter estimation, note that when the instance space X has only one element then the particular instance x can be ignored entirely. Thus the regression problem reduces to the problem of estimating the parameters of a single distribution on the outcome space Y from a sample of random outcomes y from Y, i.e. to the simpler problem of parameter estimation. Here the hypothesis is not a function but merely a single vector of parameters, and the hypothesis space $\mathcal{H}$ is the same as the decision space A.

For non-parametric density estimation, we can consider the dual case in which the outcome space Y has only one element, and hence can be ignored. Thus examples are unlabeled instances x drawn randomly from some density $p(x)$ on X. Let the decision set A be the positive real numbers and each hypothesis h in $\mathcal{H}$ be a density on X. Then, as above, information theoretic considerations suggest the the loss function $l(y,a) = l(a) = -\log a$. Again, as above, the expected loss of h is minimized when h is the true density p. Further, if p is not a member of $\mathcal{H}$, then the best hypothesis in $\mathcal{H}$, in terms of minimizing the expected loss, is the one with the smallest Kullback-Leibler divergence from the true density p [42]. When the instance space X is discrete, we let $A = (0,1)$ and each hypothesis h in $\mathcal{H}$ represent a probability distribution on X. Here we can also use the same loss function, and it has the same properties.

These examples illustrate the diversity of the learning problems that can be cast in the proposed decision theoretic framework, even under the restrictive assumptions we make in this paper, i.e. that the outcome y does not depend on the action a, and that the learner always observes both the outcome and the loss. By weakening these assumptions, we can model other types of learning as well, including *associative reinforcement learning* [11,29] and the theory of *learning automata* (with static environment) [50]. However, we will not

[4] In fact, the standard version of regression, defined as estimating the conditional mean of Y given instance x using the quadratic loss function, is actually a special case of the general version of regression defined above, where for continuous outcome spaces Y, the object is to estimate the parameters specifying the conditional density of Y given instance x, using the log likelihood loss function. To see this, assume that we represent the conditional density on Y with a Gaussian density $\hat{p}(y;\mu,\sigma) = (2\pi\sigma^2)^{-1/2} e^{-(\mu-y)^2/2\sigma^2}$, where μ is the mean and σ^2 the variance. Let the variance be fixed, independent of x, so that the estimate $\hat{p}(y;\mu,\sigma)$, of the conditional density on Y given x is completely determined by the mean μ. Thus the decision space $A \subset \Re$, and each action a in A is interpreted as specifying the mean of a Gaussian density. Substituting $\mu = a$ and evaluating $-\log \hat{p}(y;\mu,\sigma)$, the log likelihood loss is seen to be $l(y,a) = \frac{1}{2\sigma^2}(a-y)^2 + \frac{1}{2}\log(2\pi\sigma^2)$. For fixed variance σ^2, this is equivalent, for learning, to the quadratic loss $(a-y)^2$, since additive and multiplicative constants in the definition of l only rescale it without changing the value of a that minimizes its expectation.

instance x, as measured by the Kullback-Leibler divergence. It is well known that the Kullback-Leibler divergence is minimized when $\widehat{P} = P$. This is Bayes optimal hypothesis in regression.

In the regression version of our medical diagnosis situation, the definition of the log likelihood loss function depends on the interpretation of the components of the parameter vector a. If there are k possible diseases and the patient can have at most one of these, then we might have $k+1$ possible mutually exclusive disease states $y_1, \ldots, y_{k+1}$, where y_{k+1} means healthy. Hence $Y = \{y_1, \ldots, y_{k+1}\}$. Then we might specify that an action $a = (a_1, \ldots, a_{k+1})$, where the a_i are positive and sum to one, a_i being interpreted as the estimate $\widehat{P}(y_i; a)$ of the probability of disease state y_i. In this case the log likelihood loss would be $l(y_i, a) = -\log a_i$.

Often the constraints on the components of a are a nuisance, so other interpretations of a are used, e.g. that $a_i = \log \widehat{P}(y_i; a) - \log \widehat{P}(y_{k+1}; a)$ for each i, $1 \leq i \leq k+1$. In this case the $a_1, \ldots, a_k$ are arbitrary real numbers and $a_{k+1} = 0$, and hence can be ignored. Since $\widehat{P}(y_i; a) = e^{a_i} / \sum_{j=1}^{k+1} e^{a_j}$, the log likelihood loss is $l(y_i, a) = -a_i + \log \sum_{j=1}^{k+1} e^{a_j} = -a_i + \log(1 + \sum_{j=1}^{k} e^{a_j})$. This is known as the *logistic loss* [48,9]. A third interpretation would be to allow the possibility that the patient may have more than one disease, and assume, for the purposes of estimation, that diseases occur independently. Then the disease state y might be defined as a binary vector of length k, where the i^{th} bit y_i is 1 if and only if the ith disease is present. Hence $Y = \{0, 1\}^k$. Similarly, the vector a would be a vector of independent probabilities $(a_1, \ldots, a_k)$, where a_i is the estimated probability of the patient having the ith disease. In this case

$$\widehat{P}(y; a) = \prod_{i=1}^{k} a_i^{y_i} (1 - a_i)^{(1 - y_i)}$$

and the log likehood loss is

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

which we will call the *cross entropy loss*.

In the medical diagnosis example, the outcome space Y is discrete. However, in most uses of regression Y is real valued, e.g. the outcome y is the measurement of some real valued quantity, and the instance x represents the experimental conditions under which this quantity was measured. In this case regression is usually defined as estimating the conditional expectation of Y given the instance x. Thus $A \subset \Re$, and the action $a \in A$ for a given instance x consists of an estimate of the mean of the various outcomes y that would typically be observed for that instance x. It is easy to show that by using the *quadratic* loss function $l(y, a) = (a - y)^2$, the expected loss is minimized when a is the true mean, and

general, Y may be a set of strings, graphs, real vectors, etc., in which case other distance metrics or more general kinds of loss functions may be more appropriate.

1.1.3 regression

The general problem of regression has a different character from that of classification learning, but can also be addressed in our decision theoretic learning framework. To illustrate this, as a third example consider a variant of the medical diagnosis situation in which the doctor provides an estimate of the probability that the patient has each of several diseases, rather than predicting that he has one specific disease or asserting that he is healthy. (Here we assume that the actual disease state includes at most one disease.) For example, the doctor may say "Given these test results x, I would say you have disease 1 with probability 55%, disease 2 with probability 5%, and no disease at all with probability 40%." Here the doctor is actually trying to estimate the conditional distribution on disease states Y given the test results x. Her action a entails providing a vector of parameters that determine that estimated distribution, e.g. $(0.55, 0.05, 0.4)$. The decision space A is the set of all such parameter vectors.

Now let Y be an arbitrary discrete outcome space. Keeping the instance x fixed, for each parameter vector a in A and outcome y in Y let $\widehat{P}(y;a)$ denote the probability of outcome y with respect to the distribution on Y defined by the parameter vector a. Thus when we take action a on instance x, we are asserting that, given the instance x, we estimate the conditional probability of outcome y to be $\widehat{P}(y;a)$ for each outcome y in Y. Let $P(y)$ denote the actual conditional probability of outcome y, given the instance x, with respect to the unknown joint distribution on $X \times Y$. (The distributions P and $\widehat{P}$ can be replaced by densities when Y is continuous.) Let us define[2] the loss function l by setting $l(y,a) = -\log \widehat{P}(y;a)$. This is called the (negative) *log likelihood* loss function. If we define loss in this way, then the expected loss resulting from action a has a natural information theoretic interpretation[3]: it is the *Kullback-Leibler divergence* [42] (or *information gain* [64]) from the actual conditional probability distribution P to the estimated conditional distribution $\widehat{P}$, plus the entropy of P

For a given x, the entropy of the true conditional distribution P is a constant, independent of the action a. Thus choosing the action a for each instance x that minimizes the expected log likelihood loss is equivalent to choosing the action a that gives the closest estimate $\widehat{P}$ to the true conditional distribution P over possible outcomes in Y, given that

[2] We assume $\widehat{P}(y;a) > 0$ for all y in Y.

[3] The Kullback-Leibler divergence from P to $\widehat{P}$, denoted $I(P||\widehat{P})$, is defined as $\sum_{y \in Y} P(y) \log \frac{P(y)}{\widehat{P}(y;a)}$ for countable Y. The entropy of P, denoted $H(P)$, is $-\sum_{y \in Y} P(y) \log P(y)$. Thus $I(P||\widehat{P}) + H(P) = -\sum_{y \in Y} P(y) \log \widehat{P}(y;a)$, which is the expectation of the (negative) log likelihood loss. Analogous results hold for densities when the relevant quantities are finite [42].

as gain. The joint distribution on $X \times Y$ represents the probability of various races and outcomes. This joint distribution is unknown to the learner; he only has random examples $(x_1, y_1), \ldots, (x_m, y_m)$, each consisting of a race/outcome pair generated from this distribution. From these examples, the learner develops a deterministic betting strategy (hypothesis). The best hypothesis h is one that specifies a bet a for each race x that minimizes the expectation of the loss $l(y, a)$, when y is chosen randomly from the unknown conditional distribution on Y given x, which is determined by the underlying joint distribution on $X \times Y$. This (not necessarily unique) best hypothesis minimizes the expected loss on a random example (x, y). It is known as *Bayes optimal hypothesis*. The learner tries to approximate Bayes optimal hypothesis as best he can using hypotheses from a given hypothesis space $\mathcal{H}$ (e.g. "simple" or "easy to compute" hypotheses).

1.1.2 classification

As a second example, consider the problem of medical diagnosis. Here an instance x is a vector of measurements from medical tests conducted on the patient, an action a is a diagnosis of the patient's disease state, and an outcome y may be defined as the actual disease state of the patient. Here $A = Y$, i.e. the possible diagnoses are the same as the possible disease states. To specify the loss function l, we may stipulate that there is zero loss for the correct diagnosis $a = y$, but for each pair (y, a) with diagnosis a differing from disease state y there is some positive real loss $l(y, a)$, depending on the severity of the consequences of that particular misdiagnosis. Here a hypothesis is a diagnostic method, and Bayes optimal hypothesis is the one that minimizes the expected loss from misdiagnosis when examples (x, y) of test results and associated disease states occur randomly according to some unknown "natural" joint distribution.

This medical diagnosis situation is a typical example of a *classification learning* problem in the field of pattern recognition (see e.g. [20]). The problem of learning a Boolean function from noise-free examples, as investigated in the PAC model, is a special case of classification learning. Here the outcome space Y is $\{0, 1\}$ and only the instance x in an example (x, y) is drawn at random. The outcome y is $f(x)$ for some unknown Boolean *target function* f, rather than being determined stochastically. As above, the decision space A is the same as the outcome space Y, and the action a can be interpreted as a prediction of the outcome y. Hence, a hypothesis h maps from the instance space X into the outcome space Y, just as the target function does.

This same setup, where the outcome y is a function of the instance x, can be applied to any function learning problem by letting X and Y be arbitrary sets. In the general function learning problem, the loss function $l(y, a)$ usually measures the distance between the prediction a and the actual value y in some metric. In the PAC model l is the discrete metric: $l(y, a) = 0$ if $a = y$, else $l(y, a) = 1$. Thus the expected loss of the hypothesis is just the probability that it predicts incorrectly, the usual PAC notion of hypothesis error. In

1.1 Overview of the proposed framework

To extend the PAC model, we propose a more general framework based on statistical decision theory (see e.g. [14]). In this general framework we assume the learner receives randomly drawn training examples, each example consisting of an instance $x \in X$ and an outcome $y \in Y$, where X and Y are arbitrary sets called *instance* and *outcome spaces*, respectively. These examples are generated according to a joint distribution on $X \times Y$, unknown to the learner. This distribution comes from a (known) class $\mathcal{P}$ of joint distributions on $X \times Y$, representing possible "states of nature." After training, the learner will receive further random examples drawn from this same joint distribution. For each example (x, y), the learner will be shown only the instance x. Then he will be asked to choose an action a from a set of possible actions A, called the *decision space*. Following this, the outcome y will be revealed to the learner. In the case that we examine here, the outcome y depends only on the instance x and not on the action a chosen by the learner. For each action a and outcome y, the learner will suffer a loss, which is measured by a fixed real-valued *loss function* l on $Y \times A$. We assume that the loss function is known to the learner. The learner tries to choose his actions so as to minimize his loss.

Here we look at the case in which, based on the training examples, the learner develops a deterministic strategy that specifies what he believes is the appropriate action a for each instance x in X. He then uses this strategy on all future examples. Thus we look at "batch" learning rather than "incremental" or "on-line" learning [46]. The learner's strategy, which is a function from the instance space X into the decision space A, will be called a *hypothesis*. We assume that the hypothesis is chosen from a fixed *hypothesis space* $\mathcal{H}$ of functions from X into A. The goal of learning is to find a hypothesis in $\mathcal{H}$ that minimizes the expected loss, when examples are drawn at random from the unknown joint distribution on $X \times Y$.

This learning framework can be applied in a variety of situations. We now give several illustrations. For further discussion, we refer the reader to the excellent surveys of White [84], Barron [9], Devroye [19], and Vapnik [78], to which we are greatly indebted [1].

1.1.1 Betting Example

For our first example, consider the problem of learning to maximize profit (or minimize loss!) at the horse races. Here an instance x in X is a race, an action a in A consists of placing or not placing a certain bet, and an outcome y in Y is determined by the winner and the second and third place finishers. The loss $l(y, a)$ is the amount of money lost when bet a is placed and the outcome of the race is y. A negative loss is interpreted

[1] Further discussion also appears in [33]. However, there (as in previous versions of this paper) we assume that $Y = A$. In fact, we can always replace Y and A with $Y \cup A$ without loss of generality, but we now feel that the gain in economy of notation is more than offset by the confusion that results, and so have switched to the more standard notation used here.

1 Introduction

The introduction of the Probably Approximately Correct (PAC) model [75] [4] of learning from examples has done an admirable job of drawing together practitioners of machine learning with theoretically oriented computer scientists in the pursuit of a solid and useful mathematical foundation for applied machine learning work. These practitioners include both those in mainstream artificial intelligence and in neural net research. However, in attempting to address the issues that are relevant to this applied work in machine learning, a number of shortcomings of the model have cropped up repeatedly. Among these are the following:

1. The model is defined only for $\{0,1\}$-valued functions. Practitioners would like to learn functions on an instance space X that take values in an arbitrary set Y, e.g. multi-valued discrete functions, real-valued functions and vector-valued functions.

2. Some practitioners are wary of the assumption that the examples are generated from an underlying "target function", and are not satisfied with the noise models that have been proposed to weaken this assumption (e.g. [5] [70] [68]). They would like to see more general regression models investigated in which the y component in a training example $(x,y) \in X \times Y$ is randomly specified according to a conditional distribution on Y, given x. Here the general goal is to approximate this conditional distribution for each instance $x \in X$. In the computational learning theory literature, a model of this type is investigated in [38], with $Y = \{0,1\}$, and in a more general case in [85].

3. Many learning problems are unsupervised, i.e. the learner has access only to randomly drawn, unlabeled examples from an instance space X. Here learning can often be viewed as some form of approximation of the distribution that is generating these examples. This is usually called *density estimation* when the instance space X is continuous and no specific parametric form for the underlying distribution on X is assumed. It is often called *parameter estimation* when specific parametric probability models are used. One example of this in the computational learning theory literature is the recent investigation of Abe and Warmuth into the complexity of learning the parameters in a hidden Markov model [1].

The purpose of this paper is twofold. First, we propose an extension of the PAC model, based on the work of Vapnik and Chervonenkis [78] and Pollard [60,62], that addresses these and other issues. Second, we use this extension to obtain distribution-independent upper bounds on the size of the training set needed for learning with various kinds of feedforward neural networks [66] [59], a popular learning method that until now was not covered by the PAC model.

Decision Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applications

David Haussler

UCSC-CRL-91-02
September, 1989
Revised: December, 1990

haussler@saturn.ucsc.edu
Baskin Center for Computer Engineering and Information Sciences
University of California, Santa Cruz, CA 95064

Abstract: We describe a generalization of the PAC learning model that is based on statistical decision theory. In this model the learner receives randomly drawn examples, each example consisting of an instance $x \in X$ and an outcome $y \in Y$, and tries to find a hypothesis $h : X \to A$, where $h \in \mathcal{H}$, that specifies the appropriate action $a \in A$ to take for each instance x, in order to minimize the expectation of a loss $l(y, a)$. Here X, Y, and A are arbitrary sets, l is a real-valued function, and examples are generated according to an arbitrary joint distribution on $X \times Y$. Special cases include the problem of learning a function from X into Y, the problem of learning the conditional probability distribution on Y given X (regression), and the problem of learning a distribution on X (density estimation).

We give theorems on the uniform convergence of empirical loss estimates to true expected loss rates for certain hypothesis spaces $\mathcal{H}$, and show how this implies learnability with bounded sample size, disregarding computational complexity. As an application, we give distribution-independent upper bounds on the sample size needed for learning with feedforward neural networks. Our theorems use a generalized notion of VC dimension that applies to classes of real-valued functions, adapted from Pollard's work, and a notion of *capacity* and *metric dimension* for classes of functions that map into a bounded metric space.

The author gratefully acknowledges the support of ONR grant N00014-86-K-0454.