References

[1] M. Basseville and I. Nikiforov, *Detection of Abrupt Changes, Theory and Application*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[2] E. Carlstein, H. Muller, and D. Seigmund, Eds., *Change-Point Problems*. Hayward, CA: Inst. Math. Stat., 1994.

[3] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Stat.*, vol. 42, pp. 1897–1908, 1976.

[4] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *Ann. Statist.*, vol. 14, pp. 1379–1387, 1986.

[5] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, pp. 100–115, 1954.

[6] M. Pollak, "Optimal detection of a change in distribution," *Ann. Statist.*, vol. 13, pp. 206–227, 1985.

[7] M. Pollak and D. Siegmund, "Approximations to the expected sample size of certain sequential tests," *Ann. Statist.*, vol 6, pp. 1267–1282, 1975.

[8] Y. Ritov, "Decision theoretic optimality of the CUSUM procedure," *Ann. Satist.*, vol. 18, pp. 1464–1469, 1990.

[9] A. N. Shiryayev, "On optimal methods in earliest detection problems," *Theor. Probl. Appl.*, vol. 8, pp. 26–51, 1963.

[10] ——, *Optimal Stopping Rules*. New York: Springer-Verlag, 1978.

[11] A. Wald, *Sequential Analysis*. New York: Wiley, 1947.

[12] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *Ann. Math. Statist*., vol. 19, pp. 326–339, 1948.

[13] B. Yakir, "Optimal detection of a change in distribution when the observations form a Markov chain with a finite state space," in *Change-Point Problems*, E. Carlstein, H. Muller, and D. Seigmund, Eds. Hayward, CA: Inst. Math. Stat., 1994.

[14] ——, "A note on optimal detection of a change in distribution," *Ann. Statist.*, vol. 25, no. 5, pp. 2117–2126, Oct. 1997.

On the Consistency of Minimum Complexity Nonparametric Estimation

Zhiyi Chi and Stuart Geman

***Abstract*— Nonparametric estimation is usually inconsistent without some form of regularization. One way to impose regularity is through a prior measure. Barron and Cover [1], [2] have shown that complexity-based prior measures can insure consistency, at least when restricted to countable dense subsets of the infinite-dimensional parameter (i.e., function) space. Strangely, however, these results are independent of the actual complexity assignment: the same results hold under an arbitrary permutation of the match-up of complexities to functions. We will show that this phenomenon is related to the weakness of the convergence measures used. Stronger convergence can only be achieved through complexity measures that relate to the actual behavior of the functions.**

***Index Terms*—Consistency, minimum complexity estimation, minimum description length, nonparametric estimation.**

I. Introduction

Maximum-likelihood, least squares, and other estimation techniques are generally inconsistent for nonparametric (infinite-dimensional) problems. Some variety of regularization is needed. An appealing and principled approach is to base regularization on complexity: Define an encoding of the (infinite-dimensional) parameter, and adopt codelength as a penalty. Barron and Cover [1], [2] have shown how to make this work. They get consistent estimation for densities and regressions, as well as some convergence-rate bounds, by constructing complexity-based penalty terms for maximum-likelihood and least squares estimators.

Can we cite the results of Barron and Cover as an argument for complexity-based regularization (or, equivalently, for complexity-based priors)? Apparently not: The results are independent of the particular assignment of complexities. Specifically, the results are unchanged by an arbitrary permutation of the matching of complexities to parameters.

Of course there are many ways to define convergence of functions. We will show here that the surprising indifference of convergence results to complexity assignments is in fact related to the convergence measures used. Stronger convergence requires a stronger tie between the parameters (functions) and their complexity measures.

Section II is a review of some Barron and Cover results. Then some new results about consistency for nonparametric regression are presented in Section III. (Proofs are in the Appendix.) Taken together, the results of Section III establish the principle that stronger types of convergence are sensitive to the particulars of the complexity assignment. We work here with regression, but the situation is analogous in density estimation.

Our results are about consistency only. The important practical issue of relating complexity measures to *rates* of convergence remains open.

II. Complexity-Based Priors

Barron and Cover [1] have shown that the problem of estimating a density nonparametrically can be solved using a complexity-based prior by limiting the prior to a countably-dense subset of the space of densities. More specifically, given a sequence of countable sets of densities Γ_n, and numbers $L_n(q)$ for densities q in Γ_n, let $\Gamma = \cup_n \Gamma_n$. Set $L_n(q) = \infty$ for q not in Γ_n. For independent random variables $X_1, X_2, \cdots, X_n$ drawn from an unknown probability density function p, a minimum complexity density estimator $\hat{p}_n$ is defined as a density achieving the following minimization:

$$\min_{q \in \Gamma_n} \left(L_n(q) - \sum_{i=1}^{n} \log\, q(X_i) \right).$$

If we think of $L_n(q)$ as the description length of the density q, then the minimization is over total description length—accounting for both the density and the data. Barron and Cover showed that if L_n satisfies the summability condition

$$\sup_n \sum_{q \in \Gamma_n} 2^{-L_n(q)} < +\infty$$

and the growth restriction

$$\lim\, \sup_n \frac{L_n(q)}{n} = 0, \qquad \text{for every}\, q \in \Gamma \tag{1}$$

then for each measurable set S

$$\lim_{n \to \infty} \hat{P}_n(S) = P(S) \quad \text{with probability one}$$

Manuscript received December 9, 1996; revised October 20, 1997. This work was supported by the Army Research Office under Contract DAAL03-92-G-0115, the National Science Foundation, under Grant DMS-9217655, and the Office of Naval Research, under Grant N00014-96-1-0647.

The authors are with the Division of Applied Mathematics, Brown University, Providence, RI 02912 USA.

Publisher Item Identifier S 0018-9448(98)04795-6.

provided that p is in the information closure $\overline{\Gamma}$ of Γ. Here, $\hat{P}_n$ and P are the probability measures associated with the densities $\hat{p}_n$ and p, respectively, and "p is in the information closure $\overline{\Gamma}$ of Γ" means that $\inf_{q \in \Gamma} D(p \| q) = 0$, where $D(p \| q)$ is the relative entropy of p to q.

Barron and Cover also showed that if L_n satisfies a "light tail condition," i.e., if for some $0 < \alpha < 1$ and b

$$\sum_{q \in \Gamma_n} 2^{-\alpha L_n(q)} \le b, \qquad \text{for all } n \tag{2}$$

and if L_n also satisfies the growth restriction (1), then for $p \in \overline{\Gamma}$, with probability one

$$\lim_{n \to \infty} \int |p - \hat{p}_n| = 0.$$

A second paper by Barron [2] offers a minimum-complexity solution to the regression problem. Let $(X_i, Y_i)_{i=1}^n$ be independent observations drawn from the unknown joint distribution of random variables X, Y, where the support of X is in $\boldsymbol{R}^d$. Here X is the vector of explanatory variables and Y is the response variable. Functions $f(X)$ are used to predict the response. The error incurred by a prediction is measured by a distortion function $d(Y, f(X))$, the most common form being $(Y - f(X))^2$. Let h be a function which minimizes $E(d(Y, f(X)))$, which is to say that $h(x) = E(Y | X = x)$ in the squared error case. When a function f is used in place of the optimum function h the "regret" is measured by the difference between the expected distortions

$$r(f, h) = E(d(Y, f(X))) - E(d(Y, h(X))).$$

Barron defines statistical risk for a given estimator $\hat{h}_n$ to be $E(r(\hat{h}_n, h))$. Given a sequence of countable collections of functions Γ_n, and numbers $L_n(f)$, $f \in \Gamma_n$, satisfying the summability condition

$$\sup_n \sum_{f \in \Gamma_n} 2^{-L_n(f)} < \infty$$

the index of resolvability is defined as

$$R_n(h) = \min_{f \in \Gamma_n} \left(r(f, h) + \lambda \frac{1}{n} L_n(f) \right)$$

and a minimum complexity estimator is a function $\hat{h}_n \in \Gamma_n$ which achieves

$$\min_{f \in \Gamma_n} \left(\frac{1}{n} \sum_{i=1}^n d(Y_i, f(X_i)) + \lambda \frac{1}{n} L_n(f) \right).$$

Again there is a coding interpretation: if $d(Y, f(X))$ is log probability of Y given X, then $\hat{h}_n$ minimizes total description length for the model f, plus the data $Y_1, \cdots, Y_n$ given $X_1, \cdots, X_n$. Barron showed that if the support of Y and the range of each function $f(X)$ is in a known interval of length b, then with $\lambda \ge 5b^2/3 \log e$, the mean-squared error converges to zero at rate bounded by $R_n(h)$, i.e.,

$$E(r(\hat{h}_n, h)) \le O(R_n(h)). \tag{3}$$

Taken together, these results offer a general prescription for nonparametric estimation of densities and regressions. Furthermore, the connection to complexity is appealing: It is not hard to invent suitable functions $L_n(\cdot)$ by counting the bits involved in a natural encoding of Γ_n (cf. [1]). There is, however, a disturbing indifference of the results to the details of the complexity measure. For any set of permutations σ_n on Γ_n, define $L'_n(\xi) = L_n(\sigma(\xi))$ and observe that L'_n satisfies whatever conditions L_n does, and hence the same results are obtained (with the same bound on rate in (3)) using L'_n in place of L_n! In general L'_n will have no meaningful interpretation as a complexity measure.

III. What Ties Consistency to Complexity?

Suppose that X is a random variable from a probability space $(\Omega, \mathcal{F}, P)$ to $([0, 1], \mathcal{B})$. X introduces a measure P_X on $[0, 1]$ through the relation $P_X(B) = P(X^{-1}(B))$, for $B \in \mathcal{B}$. Choose a countable dense subset Γ in $L^2([0, 1], P_X)$, and define a "complexity function" L: $\Gamma \to \boldsymbol{N}$. For any random variable Y from $(\Omega, \mathcal{F}, P)$ to $(R, \mathcal{B})$ with

$$h(x) = E(Y | X = x) \in L^2([0, 1], P_X)$$

define the estimator $\hat{h}_n$ to be a function in Γ which achieves

$$\min_{f \in \Gamma} \left\{ \frac{L(f)}{n} + \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2 \right\}.$$

We will always assume that L satisfies a much stronger tail condition than (2)

$$\sum_{f \in \Gamma} e^{-\epsilon L(f)} < \infty, \qquad \text{for any } \epsilon > 0. \tag{4}$$

The first proposition demonstrates that for a weak form of convergence, consistency is essentially independent of the complexity measure:

Proposition 1: If $EY^4 < \infty$, then

$$\hat{h}_n \xrightarrow{P_X} h, \quad \text{a.s.}$$

Obviously, the proposition remains true for any permutation σ of Γ and resulting complexity function $L'(f) = L(\sigma(f))$. But, suppose we were to ask for consistency in L^2 (a.s.) in place of consistency in probability (a.s.)? Then, despite the strength of the tail condition (4), we would evidently need to pay closer attention to the complexity measure:

Proposition 2: There exists a random variable X, a countable dense subset Γ in $L^2([0, 1], P_X)$, and a function L: $\Gamma \to \boldsymbol{N}$ satisfying (4) such that for any Y with $h(x) \notin \Gamma$, the L^2 norm of $\hat{h}_n$ (in $L^2([0, 1], P_X)$) goes to $+\infty$ with probability one.

(We are focusing on the regression problem, but analogous arguments apply to probability density estimation. For example, by a construction similar to the one used for Proposition 2, the minimum complexity density estimator discussed in Barron and Cover [1] may not converge to the actual density p in the sense of Kullback–Liebler

$$\int p \log \frac{p}{\hat{p}_n} \not\to 0$$

even though the coding L satisfies the strong condition (4).)

One way to rescue consistency is to tie the complexity measure $L(f)$ more closely to f:

Proposition 3: Suppose that for every $f \in \Gamma$, $Ef^4(X) < \infty$. Assume EY^4 is finite (and hence so is $Eh^4(X)$). Construct a complexity function as follows: First, define

$$C_1(f) = (Ef^4(X) + e)e^{2Ef^2(X)}$$

and

$$C(f) = C_1(f) \log C_1(f).$$

Then, given any $L_1\colon \Gamma \to N$ which satisfies (4), let $L(f) = C(f)L_1(f)$. Then

$$\hat{h}_n \xrightarrow{L^2} h \quad \text{a.s.}$$

Proofs for the propositions are in the Appendix.

Appendix

Recall that X is a random variable defined on a probability space $(\Omega, \mathcal{F}, P)$, taking values in $([0, 1], \mathcal{B})$. P_X is defined on [0, 1] by $P_X(B) = P(X^{-1}(B))$, for $B \in \mathcal{B}$. Γ is then a countable dense subset of $L^2([0, 1], P_X)$. (Take, for example, Γ to be a countable dense set in $L^2([0, 1], dx)$; this will work for any P_X which is absolutely continuous with respect to Lebesgue measure and has bounded derivative dP_X/dx.) The complexity function $L\colon \Gamma \to \boldsymbol{N}$ is always assumed to satisfy the "strong tail condition" (4).[1] Finally, we assume that the response variable Y (a random variable on $(\Omega, \mathcal{F}, P)$) has an L^2-valued regression $h(x)$

$$h(x) = E(Y|X = x) \in L^2([0, 1], P_X).$$

The regression $h(x)$ is estimated by a function $\hat{h}_n \in \Gamma$ that achieves the minimum in

$$\min_{f\in\Gamma} \left\{ \frac{L(f)}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 \right\}.$$

We begin with Proposition 2.

Proposition 2: There exists a random variable X, a countable dense subset Γ in $L^2([0, 1], P_X)$, and a function $L\colon \Gamma \to \boldsymbol{N}$ satisfying (4) such that for any Y with $h(x) \notin \Gamma$, the L^2 norm of $\hat{h}_n$ (in $L^2([0, 1], P_X)$) goes to $+\infty$ with probability one.

Proof: Choose X so that P_X is Lebesgue measure. Fix $\Gamma = \{f_1, \cdots, f_n, \cdots\}$ dense in $L^2([0, 1], P_X)$. Let $B_1, \cdots, B_n, \cdots$ be a sequence of measurable subsets in [0, 1], each of which has positive probability, such that

$$P(\exists 1 \le i \le n,\, X_i \in B_n,\ \text{i.o. for } n) = 0.$$

This condition can be achieved, for instance, if the B's satisfy

$$\sum_{k=1}^{\infty}[1 - (1 - P_X(B_k))^k] < \infty.$$

Now for $i = 1, 2, \cdots$, define $g_i(x)$ as

$$g_i(x) = \begin{cases} f_i(x), & \text{if } x \notin B_i \\ A_i, & \text{if } x \in B_i. \end{cases}$$

We first select A_1 such that $E(g_1 - f_n)^2 > 0$ for all $n \in \boldsymbol{N}$. This can be done since there are only countably many f's while there are uncountably many choices of A_1. We then inductively select A_i such that $E(g_i - f_n)^2 > 0$, for all $n \in \boldsymbol{N}$, and $E(g_i - g_k)^2 > 0$, for $k = 1, \cdots, i - 1$. We also require of A_i that $Eg_i^2 \to +\infty$. Then $g_1, g_2, \cdots$ are distinct and none of them are in Γ. Modify Γ to include $g_1, g_2, \cdots$. Define $L\colon \Gamma \to N$ such that

$$L(f_n) > L(g_n)$$

and

$$\sum_{f\in\Gamma} e^{-\epsilon L(f)} < \infty, \qquad \text{for any } \epsilon > 0.$$

Now given Y, with

$$h(x) = E(Y|X = x) \in L^2([0, 1], P_X)$$

and $h(x) \notin \Gamma$, the set of ω which satisfies

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 \to E(h(X) - f(X))^2 + E(Y - h(X))^2, \qquad \forall f \in \Gamma$$

and

$$X_i(\omega) \notin B_n,\ \forall 1 \le i \le n,\ \forall \text{large } n$$

is of probability one. For any ω in this set, let

$$I_n(\omega) = \arg\min_{k} \left\{ \frac{L(f_k)}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_k(X_i))^2 \right\}.$$

Then since $h \notin \Gamma$, $I_n(\omega) \to \infty$ as $n \to \infty$. For large n, $X_i(\omega) \notin B_{I_n(\omega)}$ for all $1 \le i \le I_n(\omega)$, and hence

$$g_{I_n(\omega)}(X_i(\omega)) = f_{I_n(\omega)}(X_i(\omega))\, \forall 1 \le i \le I_n(\omega).$$

Therefore, for large n

$$\frac{L(g_{I_n})}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - g_{I_n}(X_i))^2 < \frac{L(f_{I_n})}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_{I_n}(X_i))^2.$$

Consequently, with probability one, for large n

$$\hat{h}_n = \arg\min_{f\in\Gamma} \left\{ \frac{L(f)}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 \right\} \in \{g_1, g_2, \cdots\}.$$

Since $E(g_i^2) \to \infty$, this completes the proof. □

Remark: As mentioned in Section III, the same argument can be used to show that the minimum complexity estimator $\hat{p}_n$ in [1] may not converge to the true density p, in the sense that

$$\int p \log \frac{p}{\hat{p}_n} \not\to 0.$$

The proof of Proposition 1 is based on the following three lemmas.

Lemma 1: Fix $\epsilon > 0$. Let $Z_1, Z_2, \cdots, Z_n$ be a sequence of independent and identically distributed (i.i.d.) random variables satisfying

a) $Z_1 \ge 0$;
b) $EZ_1^2 < \infty$.

Then if

$$K \ge (\operatorname{Var}(Z_1) + \epsilon^2)e^{EZ_1} \quad \text{and} \quad \frac{\epsilon}{K} < 1$$

then

$$P\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - EZ_1) \le -\epsilon\right) \le \left(1 - \frac{\epsilon^2}{2K}\right)^n.$$

Proof: For any $t \in (0, 1]$

$$P\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - EZ_1) \le -\epsilon\right) \le \left(Ee^{t(-Z_1+EZ_1-\epsilon)}\right)^n.$$

[1] For example, choose $a(\cdot)$ strictly positive such that $\sum_f a(f) < \infty$. If $F(x)$ is any strictly positive function satisfying $F(x)/x \to \infty$ as $x \to \infty$, then $L(f) = F(-\log a(f))$ satisfies (4).

Let $\phi(t) = Ee^{t(-Z_1+EZ_1-\epsilon)}$, then

$$\phi(0) = 1 \qquad \phi'(0) = -\epsilon$$

and

$$\begin{aligned}\phi''(t) &= E((Z_1 - EZ_1 + \epsilon)^2 e^{t(-Z_1+EZ_1-\epsilon)})\\ &\le E(Z_1 - EZ_1 + \epsilon)^2 e^{tEZ_1} \le K, \qquad \text{for } t \in (0, 1].\end{aligned}$$

Hence

$$\phi'(t) \le -\epsilon + Kt, \qquad \text{for } t \in (0, 1]$$

and

$$\phi(t) \le 1 - \epsilon t + \tfrac{1}{2}Kt^2, \qquad \text{for } t \in (0, 1].$$

Take $t = \epsilon/K < 1$, which is the minimizer of $1 - \epsilon t + Kt^2/2$. Then

$$P\left(\frac{1}{n}\sum_{i=1}^{n} Z_i < EZ_1 - \epsilon\right) \le \left(1 - \frac{\epsilon^2}{2K}\right)^n. \qquad \square$$

Lemma 2: Suppose $EY^4 < \infty$. Let

$$h(x) = E(Y|X = x) \in L^2([0, 1], P_X).$$

Assume Γ is a countable dense subset of

$$\{f \in L^2([0, 1], P_X)\colon |f(x)| \le M\}$$

and $L\colon \Gamma \to \boldsymbol{N}$ satisfies condition (4). Then given $0 < \epsilon < 1$, with probability one, for sufficiently large n and all $f \in \Gamma$ with $E(f - h_M)^2 \ge 3\epsilon$

$$\frac{1}{n}\sum_{i=1}^{n}(h_M(X_i) - Y_i)^2 + \epsilon < \frac{L(f)}{n} + \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2 \quad (5)$$

where for any function f

$$f_M(x) = \begin{cases} f(x), & \text{if } |f(x)| \le M\\ \text{sign}\,(f(x)) \cdot M, & \text{otherwise.}\end{cases} \quad (6)$$

Proof: We shall first give the idea of the proof. Assume $|h| < M$. With probability one, when n is sufficiently large

$$\frac{1}{n}\sum_{i=1}^{n}(h(X_i) - Y_i)^2 + \epsilon$$

is bounded by $E(h(X) - Y)^2 + 2\epsilon$. We then get a stronger inequality

$$E(h(X) - Y)^2 + 2\epsilon \le \frac{L(f)}{n} + \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2.$$

The left-hand side equals

$$E(f(X) - Y)^2 - E(f(X) - h(X))^2 + 2\epsilon \le E(f(X) - Y)^2 - \epsilon.$$

Hence we can prove the lemma by showing

$$\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2 - E(f(X) - Y)^2 > -\epsilon - \frac{L(f)}{n}$$

is true with probability one, for sufficiently large n and all $f \in \Gamma$. By Lemma 1, for each fixed n and $f \in \Gamma$, the probability that this inequality does not hold is bounded by

$$\left(1 - \frac{(\epsilon + L(f)/n)}{K}\right)^n \le \left(1 - \frac{\epsilon^2}{K}\right)^n \left(1 - \frac{\epsilon L(f)/n}{K}\right)^n$$

where K is a large number independent of n. Because $1 - x < e^{-x}$ for all $x > 0$, the above probability is then bounded by

$$\left(1 - \frac{\epsilon^2}{K}\right)^n e^{-\epsilon L(f)/K}.$$

Summing over all $f \in \Gamma$, we see that the probability that (5) is not true is exponentially small. A Borel–Cantelli argument then finishes the proof.

We turn now to the details of the proof. Define

$$B(h_M) = \{f \in \Gamma\colon E(f - h_M)^2 \ge 3\epsilon\}. \quad (7)$$

For $f \in \Gamma$, define

$$T_{f,n}(h_M) = \left\{\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2 + \frac{L(f)}{n} \le \frac{1}{n}\sum_{i=1}^{n}(h_M(X_i) - Y_i)^2 + \epsilon\right\} \quad (8)$$

$$V_n(h_M) = \bigcup_{f \in B(h_M)} T_{f,n}. \quad (9)$$

Write

$$R_n(h_M) = \left\{\left|\frac{1}{n}\sum_{i=1}^{n}(h_M(X_i) - Y_i)^2 - E(h_M(X) - Y)^2\right| < \epsilon\right\} \quad (10)$$

$$R(h_M) = \liminf_{n\to\infty} R_n(h_M). \quad (11)$$

Henceforth, we will simplify the notation by writing B instead of $B(h_M)$, $T_{f,n}$ instead of $T_{f,n}(h_M)$, and so on. By the strong law of large numbers, $P(R) = 1$. Next show that $\sum_n P(V_n \cap R_n) < \infty$. If this is true, then by the Borel–Cantelli lemma

$$P(\limsup_{n\to\infty} V_n) = P(\limsup_{n\to\infty} V_n \cap R) \le P(\limsup_{n\to\infty}(V_n \cap R_n)) = 0$$

which is what needs to be proved.

For $\omega \in R_n$ and $f \in B$

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}(h_M(X_i) - Y_i)^2 + \epsilon - E(f(X) - Y)^2\\ &\qquad \le 2\epsilon + E(h_M(X) - Y)^2 - E(f(X) - Y)^2.\end{aligned}$$

Clearly,

$$E(Y - f(X))^2 = E(Y - h(X))^2 + E(h(X) - f(X))^2.$$

Since $|f| \le M$, $|h - f| = |h - h_M| + |h_M - f|$

$$\begin{aligned}E(Y - f(X))^2 &\ge E(Y - h(X))^2 + E(h(X) - h_M(X))^2\\ &\quad + E(h_M(X) - f(X))^2\\ &= E(Y - h_M(X))^2 + E(h_M(X) - f(X))^2\\ &\ge E(Y - h_M(X))^2 + 3\epsilon.\end{aligned}$$

Hence

$$\frac{1}{n}\sum_{i=1}^{n}(h_M(X_i) - Y_i)^2 + \epsilon - E(f(X) - Y)^2 \le -\epsilon.$$

Suppose $f \in B$ and $R_n \cap T_{f,n} \neq \emptyset$. For any $\omega \in R_n \cap T_{f,n}$, by the above inequality

$$\frac{1}{n}\sum_{i=1}^{n}(f(X_i)-Y_i)^2 - E(f(X)-Y)^2 \le -\epsilon - \frac{L(f)}{n} = -\delta_{f,n}.$$

Furthermore,

$$\frac{L(f)}{n} \le \frac{1}{n}\sum_{i=1}^{n}(h_M(X_i)-Y_i)^2 + \epsilon$$

and hence

$$\begin{aligned}\delta_{f,n} &\le 2\epsilon + \frac{1}{n}\sum_{i=1}^{n}(h_M(X_i)-Y_i)^2\\ &\le 3 + E(h_M(X)-Y)^2 = H. \qquad (12)\end{aligned}$$

Fix K such that

$$K \ge (E(M+|Y|)^4 + H^2)e^{E(M+|Y|)^2}.$$

Now for any $f \in B$ with $R_n \cap T_{f,n} \neq \emptyset$, it is easy to check

$$(\mathrm{Var}\,((f(X)-Y)^2) + \delta_{f,n}^2)e^{E(f(X)-Y)^2} \le K \quad \text{and} \quad \delta_{f,n} < K.$$

Then by Lemma 1, for any $f \in B$ with $R_n \cap T_{f,n} \neq \emptyset$

$$\begin{aligned}&P(R_n \cap T_{f,n})\\ &\quad\le P\left(\frac{1}{n}\sum_{i=1}^{n}(f(X_i)-Y_i)^2 - E(f(X)-Y)^2 \le -\delta_{f,n}\right)\\ &\quad\le \left(1 - \frac{(L(f)/n+\epsilon)^2}{2K}\right)^n\\ &\quad\le \left(1-\frac{\epsilon^2}{2K}\right)^n\left(1-\frac{\epsilon L(f)/n}{K}\right)^n.\end{aligned}$$

Since

$$\frac{\epsilon L(f)/n}{K} < \frac{\epsilon\delta_{f,n}}{K} < 1$$

and $1 - x < e^{-x}$, for all $0 < x < 1$, we get $P(R_n \cap T_{f,n})$ is bounded by

$$\left(1-\frac{\epsilon^2}{2K}\right)^n \exp\left(-\frac{\epsilon L(f)}{K}\right).$$

Therefore,

$$\begin{aligned}P(R_n \cap V_n) &\le \sum_{f\in B} P(R_n \cap T_{f,n})\\ &\le \left(1-\frac{\epsilon^2}{2K}\right)^n \sum_{f\in\Gamma}\exp\left(-\frac{\epsilon L(f)}{K}\right)\end{aligned}$$

and by the strong tail condition (4), $\sum \exp(-\epsilon L(f)/K) < \infty$. Since K is independent of n, $P(R_n \cap V_n)$ is exponentially small and $\sum P(R_n \cap V_n)$ converges. □

Lemma 3: Let μ be a finite measure, and let f and f_n, $n = 1, 2, \cdots$, be measurable functions. If $f < \infty$, μ-a.s., and if

$$\liminf_{M\to\infty}\ \limsup_{n\to\infty} E(f_{n,M} - f_M)^2 = 0$$

then $f_n \xrightarrow{\mu} f$.

Proof: Suppose $M_n \to \infty$ is a sequence such that

$$\lim_{k\to\infty}\ \limsup_{n\to\infty} E(f_{n,M_k} - f_{M_k})^2 = 0.$$

Fix $\epsilon > 0$ and $M > 0$. Then

$$\begin{aligned}\mu(\{|f_n - f| > \epsilon\}) &\le \mu(\{|f| \ge M_k - \epsilon\})\\ &\quad + \mu(\{|f| < M_k - \epsilon,\ |f_{n,M_k} - f_{M_k}| > \epsilon\})\\ &\le \mu(\{|f| \ge M_k - \epsilon\}) + \frac{1}{\epsilon^2}E(f_{n,M_k} - f_{M_k})^2.\end{aligned}$$

Let $n \to \infty$ and then $k \to \infty$ to complete the proof. □

Proposition 1: If $EY^4 < \infty$, then

$$\hat{h}_n \xrightarrow{P_X} h, \quad \text{a.s.}$$

Proof: The idea is to choose $M_k \to \infty$ and then truncate the functions in Γ as in (6). Then by Lemma 2, we will get $E(\hat{h}_{n,M_k} - h_{M_k})^2 \to 0$, where $\hat{h}_{n,M_k}$ is the truncated $\hat{h}_n$, and h_{M_k} is the truncated h. We then use Lemma 3 to get $\hat{h}_n \overset{P_X}{\to} h$.

Filling in the details, given $\epsilon > 0$, there is $M = M(\epsilon) > 0$ such that $E(h - h_M)^2 < \epsilon$ and

$$\int_{|Y|>M}(|Y| + M)^2 \le 4\int_{|Y|>M}|Y|^2 < \epsilon.$$

With probability one, when n is sufficiently large

$$\frac{L(\hat{h}_n)}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{h}_n(X_i))^2 < \frac{1}{n}\sum_{i=1}^{n}(Y_i - h_M(X_i))^2 + \epsilon.$$

Consider

$$\frac{L(\hat{h}_n)}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{h}_{n,M}(X_i))^2.$$

Observe that $|Y_i - \hat{h}_{n,M}(X_i)| > |Y_i - \hat{h}_n(X_i)|$ implies $|Y_i| > M$. Hence

$$\begin{aligned}&\frac{L(\hat{h}_n)}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{h}_{n,M}(X_i))^2\\ &\qquad\le \frac{L(\hat{h}_n)}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{h}_n(X_i))^2\\ &\qquad\quad + \frac{1}{n}\sum_{i=1}^{n}(|Y_i| + M)^2 \cdot I_{|Y_i|>M}.\end{aligned}$$

With probability one, for sufficiently large n

$$\frac{1}{n}\sum_{i=1}^{n}(|Y_i| + M)^2 \cdot I_{|Y_i|>M} \le \int_{|Y|>M}(|Y| + M)^2 + \epsilon < 2\epsilon$$

and, therefore, for large n

$$\frac{L(\hat{h}_n)}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{h}_{n,M}(X_i))^2 \le \frac{1}{n}\sum_{i=1}^{n}(Y_i - h_M(X_i))^2 + 3\epsilon.$$

Let $\Gamma_M = \{f_M: f \in \Gamma\} \cup \{h_M\}$, which is dense in

$$L^2([0, 1], P_X) \cap \{\|f\|_\infty \le M\}.$$

Define L': $\Gamma_M \to \boldsymbol{N}$ as

$$L'(\pi) = \min\{L(f): f_M = \pi, f \in \Gamma\}.$$

Then with probability one, for large n

$$\frac{L'(\hat{h}_{n,M})}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{n,M}(X_i))^2 \le \frac{1}{n}\sum_{i=1}^{n}(Y_i - h_M(X_i))^2 + 3\epsilon.$$

L' satisfies the strong tail condition (4). According to Lemma 2, with probability one, for sufficiently large n

$$E(\hat{h}_{n,M} - h_M)^2 \le 9\epsilon.$$

Let $S(\epsilon)$ be the subset of points in Ω such that the above relation holds, i.e.,

$$S(\epsilon) = \liminf_{n\to\infty}\left\{\omega: E(\hat{h}_{n,M} - h_M)^2 \le 9\epsilon\right\}.$$

Choose a sequence $\epsilon_n \to 0$, and let $M_n = M(\epsilon_n)$ and $S_n = S(\epsilon_n)$. Then on $S = \cap S_n$, which has probability one

$$\lim_{k\to\infty}\sup \lim_{n\to\infty} E(\hat{h}_{n,M_k} - h_{M_k})^2 = 0.$$

By Lemma 3, for any $\omega \in S$, $\hat{h}_n \xrightarrow{P_X} h$, which completes the proof.□

Proposition 3: Suppose that for every $f \in \Gamma$, $Ef^4(X) < \infty$. Assume EY^4 is finite (and hence so is $Eh^4(X)$). Construct a complexity function as follows: First, define

$$C_1(f) = (Ef^4(X) + e)e^{2Ef^2(X)}$$

and

$$C(f) = C_1(f) \log C_1(f).$$

Then, given any L_1: $\Gamma \to N$ which satisfies (4), let $L(f) = C(f)L_1(f)$. Then

$$\hat{h}_n \xrightarrow{L^2} h \quad \text{a.s.}$$

Proof: We will follow closely the proof and the notation of Lemma 2. As in Lemma 2, we need to show that $P(\limsup V_n) = 0$. Fixing a number $D = D(Y, h, \epsilon)$, which will be determined later, we first decompose V_n as

$$V_n = \bigcup_{f\in B} T_{f,n} = \bigcup_{f\in B,\, L_1(f)\ge D} T_{f,n} \cup \bigcup_{f\in B,\, L_1(f)<D} T_{f,n} = V_n' \cup V_n''.$$

Since there are only finitely many f with $L_1(f) < D$, by the strong law of large numbers, $P(\limsup V_n'') = 0$. Thus in order to get $P(\limsup V_n) = 0$, we need only show that $P(\limsup V_n') = 0$. Similar to Lemma 2, it is enough to check

$$\sum_n P(V_n' \cap R_n) < \infty.$$

Derive again the constant H, as in (12). Then for each $f \in \Gamma$, define

$$K(f) = (\mathrm{Var}\,((f(X) - Y)^2) + H^2)e^{E(f(X)-Y)^2} > e.$$

Then for any $f \in B$ with $R_n \cap T_{f,n} \ne \emptyset$, as in the proof of Lemma 2

$$P(R_n \cap T_{f,n}) \le \left(1 - \frac{\epsilon^2}{2K(f)}\right)^n \exp\left(-\frac{\epsilon C(f)L_1(f)}{K(f)}\right).$$

Hence

$$\begin{aligned}\sum_{n=1}^{\infty} P(R_n \cap V_n') &\le \sum_{L_1(f)\ge D}\sum_{n=1}^{\infty} P(R_n \cap T_{f,n}) \\ &\le \sum_{L_1(f)\ge D} \frac{2K(f)}{\epsilon^2} \exp\left(-\frac{\epsilon C(f)L_1(f)}{K(f)}\right) \\ &= \frac{2}{\epsilon^2}\sum_{L_1(f)\ge D} \exp\,(L_1(f)J(f, \epsilon))\end{aligned}$$

where

$$J(f, \epsilon) = -\frac{\epsilon C(f)}{K(f)} + \frac{\log K(f)}{L_1(f)}.$$

It is easy to see that there is a constant $c = c(Y, h) > 0$, such that $C(f) \ge cK(f) \log K(f) > 0$. Now choose $D = D(Y, h, \epsilon)$ such that $\epsilon cD \ge 2$. Then for $L_1(f) \ge D$

$$\frac{\log K(f)}{L_1(f)} \le \frac{\epsilon C(f)}{2K(f)}.$$

Since $K(f) > e$

$$J(f, \epsilon) \le -\frac{\epsilon C(f)}{2K(f)} \le -\frac{\epsilon C(f)}{2K(f) \log K(f)} \le -\frac{\epsilon c}{2}.$$

So

$$\sum_{n=1}^{\infty} P(R_n \cap V_n') \le \frac{2}{\epsilon^2}\sum_{f\in\Gamma} e^{-\epsilon c L_1(f)/2} < \infty.$$

Similar to Lemma 3, we can now conclude that for any $0 < \epsilon < 1$, the set

$$S(\epsilon) = \left\{\omega: E(\hat{h}_n - h)^2 < 3\epsilon, \text{ for sufficiently large } n\right\}$$

has probability one. Finally, then, for $\omega \in \cap_{k=1}^{\infty} S(k^{-1})$

$$E(\hat{h}_n - h)^2 \to 0 \text{ as } n \to \infty. \qquad \square$$

References

[1] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1034–1054, July 1991.

[2] A. R. Barron, "Complexity regularization with application to artificial neural networks," in *Non-parametric Functional Estimation and Related Topics*, G. Roussas, Ed. Boston, MS: Kluwer.