April, 1997

LIDS-P 2387

Research Supported By:

Presidential Young Investigator Award
Matching funds Draper Laboratory
ARO grant DAAL-03-92-G-0115
NSF grant DDM-9158118

Estimation of Time-Varying Parameters in Statistical Models; an Optimization Approach

Bertsimas, D.
Gamarnik, D.
Tsitsiklis, J.N.

Estimation of Time-Varying Parameters in Statistical Models; an Optimization Approach

Dimitris Bertsimas *
Sloan School of Management and
Operations Research Center, MIT
Cambridge, MA 02139
dbertsim@aris.mit.edu

David Gamarnik †
Operations Research Center, MIT
Cambridge, MA 02139
dpg6138@mit.edu

John N.Tsitsiklis ‡
Laboratory for Information and
Decision Sciences
and Operations Research Center, MIT
Cambridge, MA 02139
jnt@mit.edu

Abstract

We propose a convex optimization approach to solving the nonparametric regression estimation problem when the underlying regression function is Lipschitz continuous. This approach is based on minimizing the sum of empirical squared errors, subject to the constraints implied by the Lipschitz continuity. The resulting optimization problem has a convex objective function and linear constraints, and as a result, is efficiently solvable. The estimating function, computed by this technique, is proven to converge to the underlying regression function uniformly and almost surely, when the sample size grows to infinity, thus providing a very strong form of consistency.

We also propose a convex optimization approach to the maximum likelihood estimation of unknown parameters in statistical models where parameters depend continuously on some observable input variables. For a number of classical distributional forms, the objective function in the underlying optimization problem is convex, and the constraints are linear. These problems are therefore also efficiently solvable.

1 Introduction

Nonlinear regression is the process of building a model of the form

$$Y = f(X) + \psi, \tag{1}$$

where X, Y are observable random variables and ψ is a zero-mean non-observable random variable. Thus, $E[Y|X] = f(X)$. The main problem of nonlinear regression analysis is to estimate a function f based on a sequence of observations $(X_1, Y_1), \dots, (X_n, Y_n)$. In one particular instance, we may think of variable X_i as the time t_i at which we observed Y_i. That is, at times $t_1 < t_2 < \cdots < t_n$ we observed $Y_1, Y_2, \dots, Y_n$, and the problem is to compute a time varying mean value $E[Y(t)]$ of Y as a function of time t on the interval $[t_1, t_n]$. However, this paper also considers the case where the dimension of X is larger than one.

There are two mainstream approaches to the problem. The first is parametric estimation, where some specific form of the function f is assumed (for example, f is a polynomial) and unknown parameters (for example the coefficients of the polynomial) are estimated.

The second approach is nonparametric regression. This approach usually assumes only qualitative properties of the function f, like differentiability or square integrability. Among the various nonparametric regression techniques, the two best known and most understood are kernel regression and smoothing splines (see [2] for a systematic treatment).

Consistency (convergence of the estimate to the true function f as sample size goes to infinity) is known to hold for both of these techniques. Also for the case of a one-dimensional input vector X, the decay rates of the magnitudes of expected errors are known to be of order $O(\frac{1}{n^{4/5}})$ for kernel regression and $O(\frac{1}{n^{m/m+1}})$ for smoothing splines, where m stands for the number of continuous derivatives existing for the function f.

In this paper, we show how convex optimization techniques can be used in nonparametric regression, when the underlying function to be estimated is Lipschitz continuous. The idea is to minimize the sum of the empirical squared errors subject to constraints implied by Lipschitz continuity. This method is therefore very close in spirit to the smoothing splines approach which is built on minimizing the sum of squared errors and penalizing large magnitude of second or higher order derivatives. But, unlike smoothing splines, our technique

* Research partially supported by a Presidential Young Investigator Award DDM-9158118 with matching funds from Draper Laboratory.

† Research partially supported by the ARO under grant DAAL-03-92-G-0115 and by the NSF under grant DDM-9158118.

‡ Research partially supported by the ARO under grant DAAL-03-92-G-0115.

does not require differentiability of the regression function and, on the other hand, enforces the Lipschitz continuity constraint, so that the resulting approximation is a Lipschitz continuous function.

The contributions of the paper are summarized as follows:

1. We propose a convex optimization approach to the nonlinear regression problem. Given an observed sequence of inputs $X_1, X_2, \ldots, X_n$, and outputs $Y_1, Y_2, \ldots, Y_n$, we compute a Lipschitz continuous estimating function $\hat{f}^n \equiv \hat{f}(X_1, Y_1, \ldots, X_n, Y_n)$ with a specified Lipschitz constant K. Thus our method is expected to work well when the underlying regression function f is itself Lipschitz continuous and the constant can be guessed within a reasonable range (see simulations results in Section 5 and Theorem 6.1 in Section 6).

2. In Section 3, we outline the convex optimization approach to the maximum likelihood estimation of unknown parameters in dynamic statistical models. It is a modification of the classical maximum likelihood approach, but to models with parameters depending continuously on some observable input variables.

3. Our main theoretical results are contained in Section 6. For the case of bounded random variables X and Y, we establish a very strong mode of convergence of the estimating function $\hat{f}^n$ to the true function f, where n is the sample size. In particular, we show that $\hat{f}^n$ converges to f *uniformly and almost surely,* as n goes to infinity. We also establish that the tail of the distribution of the uniform distance $||\hat{f}^n - f||_\infty$ decays exponentially fast. Similar results exist for kernel regression estimation [3], but do not exist, to the best of our knowledge, for smoothing splines estimators.

   Uniform convergence coupled with the exponential bound on the tail of the distribution of $||\hat{f}^n - f||_\infty$ enables us to build confidence intervals around $\hat{f}^n$. However, the constants in our tail distribution estimations might be too large for practical purposes.

2 A nonlinear regression model

In this section, we demonstrate how convex optimization algorithms can be used for nonlinear regression analysis.

The objective is to find an estimator $\hat{f}$ of the true function f in model (1) based on the sequence of observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$:

$$Y_i = f(X_i) + \psi_i, \qquad i = 1, 2, \ldots, n.$$

We denote by $\mathcal{X} \subset \Re^m$ and $\mathcal{Y} \subset \Re$ the ranges of the vector X and the random variable Y. Let also $||\cdot||$ denote the Euclidean norm in the vector space $\Re^m$.

We propose the following two step algorithm.

Regression algorithm

Step 1. Choose a constant K and solve the following constrained optimization problem in the variables $\hat{f}_1, \ldots, \hat{f}^n$:

$$\begin{aligned} &\text{minimize} \sum_{i=1}^{n} (Y_i - \hat{f}_i)^2 \\ &\text{subject to} \qquad\qquad\qquad\qquad (2) \\ &|\hat{f}_i - \hat{f}_j| \leq K||X_i - X_j||, \qquad i, j = 1, 2, \ldots, n. \end{aligned}$$

This step gives the prediction of the output $\hat{f}_i \equiv \hat{f}(X_i)$, $i = 1, 2, \ldots, n$ at the inputs $X_1, X_2, \ldots, X_n$.

Step 2. In this step, we extrapolate the values $\hat{f}_1, \ldots, \hat{f}^n$ obained in Step 1, to a Lipschtitz continuous function $\hat{f} : \mathcal{X} \to \Re$ with the constant K as follows: for any $x \in \mathcal{X}$, let

$$\hat{f}(x) = \max_{1 \leq i \leq n} \{\hat{f}_i - K||x - X_i||\}.$$

The following proposition justifies Step 2 of the above algorithm.

Proposition 2.1 *The function $\hat{f}$ defined above is a Lipschitz continuous function with Lipschitz constant K. It satisfies*

$$\hat{f}(X_i) = \hat{f}_i, \qquad i = 1, 2, \ldots, n.$$

Proof: Let $x_1, x_2 \in \mathcal{X}$. Let $i = \text{argmax}_{1 \leq j \leq n}\{\hat{f}_j - K||x_1 - X_j||\}$ i.e. $\hat{f}(x_1) = \hat{f}_i - K||x_1 - X_i||$. Moreover, by the definition of $\hat{f}(x_2)$, $\hat{f}(x_2) \geq \hat{f}_i - K||x_2 - X_i||$. Therefore,

$$\hat{f}(x_1) - \hat{f}(x_2) \leq \hat{f}_i - K||x_1 - X_i|| - (\hat{f}_i - K||x_2 - X_i||) =$$

$$= K||x_2 - X_i|| - K||x_1 - X_i|| \leq K||x_2 - x_1||.$$

By a symmetric argument, we obtain

$$\hat{f}(x_2) - \hat{f}(x_1) \leq K||x_2 - x_1||.$$

For $x = X_i$, we have $\hat{f}_i - K||x - X_i|| = \hat{f}_i$. For all $j \neq i$, constraint (2) guarantees $\hat{f}_j - K||x - X_j|| \leq \hat{f}_i$. It follows that $\hat{f}(X_i) = \hat{f}_i$. □

In Step 2, we could take instead

$$\hat{f}(x) = \min_{1 \leq i \leq n} \{\hat{f}_i + K||x - X_i||\},$$

or

$$\hat{f}(x) = \frac{1}{2} \max_{1 \leq i \leq n} \{\hat{f}_i - K||x - X_i||\} + \frac{1}{2} \min_{1 \leq i \leq n} \{\hat{f}_i + K||x - X_i||\}.$$

Proposition 2.1 holds for the both of these constructions.

Interesting special cases of model (1) include dynamic models. Suppose that $X_1, \ldots, X_n$ are times at which measurements $Y_1, \ldots, Y_n$ were observed. That is, at times $t_1 < t_2 < \cdots < t_n$ we observe $Y_1, \ldots, Y_n$. To estimate the time varying expectation of the random variable Y within the time interval $[t_1, t_n]$, we modify the two steps of the regression algorithm as follows:

Step 1'. Solve the following optimization problem in the variables $\hat{f}_1, \ldots, \hat{f}_n$

$$\begin{aligned} &\text{minimize} \sum_{i=1}^{n} (Y_i - \hat{f}_i)^2 \\ &\text{subject to} \qquad\qquad\qquad\qquad\qquad (3) \\ &|\hat{f}_{i+1} - \hat{f}_i| \leq K(t_{i+1} - t_i), \qquad i = 1, 2, \ldots, n-1 \end{aligned}$$

Step 2'. The extrapolation step can be performed in the following way. For any t, with $t_i \leq t < t_{i+1}$, let

$$\mu = \frac{t - t_i}{t_{i+1} - t_i},$$

and set

$$\hat{f}(t) = (1 - \mu)\hat{f}(t_i) + \mu \hat{f}(t_{i+1}).$$

It is easy to see that the resulting function $\hat{f}$ defined on the interval $[t_1, t_n]$ is Lipschitz continuous with constant K.

Remarks:

1. The motivation of the proposed algorithm is to try to minimize the sum of the empirical squared errors between the estimated function value $\hat{f}_i$ at point X_i and the observed one Y_i, in such a way that the estimations $\hat{f}_1, \ldots, \hat{f}^n$ satisfy the Lipschitz continuity condition.

2. The choice of the constant K is an important part of the setup. It turns out that for a successful approximation, it suffices to take $K \geq K_0$, where K_0 is the true Lipschitz constant of the unknown function f (see Section 6).

3. If the noise terms $\psi_1, \ldots, \psi_n$, are i.i.d., then this approach also yields an estimate of the variance of the noise ψ:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \hat{f}_i\right)^2$$

4. Optimization problems (2) or (3) can be solved efficiently, since the objective function is quadratic (convex) and all the constraints are linear, (see [4].)

5. Setting $K = 0$, yields a usual sample average:

$$\hat{f}_1 = \cdots = \hat{f}^n = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

6. If the noise terms $\psi_1, \ldots, \psi_n$, are identically zero, then the estimating function $\hat{f}$ coincides with the true function f on the observed input values:

$$\hat{f}_i = f(X_i), \qquad i = 1, 2, \ldots, n.$$

   This compares favorably with the kernel regression technique, where due to the selected positive bandwidth, the estimating function is not equal to the true function even if the noise is zero. Thus, our method is robust with respect to small noise levels.

It is clear that we cannot expect the pointwise unbiasedness condition $E[\hat{f}(x)] = f(x)$ to hold universally for all $x \in \mathcal{X}$. However, the estimator produced by our method is unbiased in an *average* sense as the following theorem shows.

Theorem 2.1 *Let estimators $\hat{f}_i$ be obtained from the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ as outlined in Step 1 of the regression algorithm. Then,*

$$E\left[\frac{1}{n} \sum_{i=1}^{n} \hat{f}_i \;\middle|\; X_1, \ldots, X_n\right] = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Proof: Let the sequence $\hat{f}_1, \ldots, \hat{f}_n$ be obtained using Step 1 of the regression algorithm. Observe that the sequence $\hat{f}_i + c$, $i = 1, 2, \ldots, n$, also satisfies the constraints in (2), for any $c \in \mathcal{R}$. That is, all the costs

$$\sum_{i=1}^{n} (Y_i - \hat{f}_i - c)^2, \qquad c \in \mathcal{R}$$

are achievable. For fixed $Y_i, \hat{f}_i$, $i = 1, 2, \ldots, n$ the minimal cost is achieved for

$$c^* = \sum_{i=1}^{n} (Y_i - \hat{f}_i),$$

However, we have that $\sum_{i=1}^{n} (Y_i - \hat{f}_i)^2$ is a minimal achievable cost. Therefore

$$c^* = \sum_{i=1}^{n} (Y_i - \hat{f}_i) = 0.$$

or

$$\frac{1}{n} \sum_{i=1}^{n} \hat{f}_i = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

It follows that

$$\begin{aligned} E\left[\frac{1}{n} \sum_{i=1}^{n} \hat{f}_i \;\middle|\; X_1, \ldots, X_n\right] &= E\left[\frac{1}{n} \sum_{i=1}^{n} Y_i \;\middle|\; X_1, \ldots, X_n\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} f(X_i), \end{aligned}$$

which completes the proof. □

3 A general dynamic statistical model

We now propose a convex optimization approach for maximum likelihood estimation of parameters, that depend on some observable input variable.

Given a sequence of input-output random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$, suppose the random variables $Y_i, i = 1, 2, \ldots, n$, are distributed according to some *known* probability density function $\phi(\lambda)$, which depends on some parameter λ. This parameter is *unknown* and is a Lipschitz continuous function $\lambda : \mathcal{X} \longrightarrow \Re$ (with unknown constant K_0) of the input variable X.

In particular, Y_i has a probability density function $\phi(\lambda(X_i), Y_i),\ i = 1, 2, \ldots, n$, where $\phi(\cdot)$ is a known function, and $\lambda(\cdot)$ is unknown. The objective is to estimate the true parameter function λ based on the sequence of i.i.d. observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. As a solution we propose the following algorithm

Dynamic Maximum Likelihood Estimation Algorithm (DMLE algorithm)

Step 1. Solve the following optimization problem in the variables $\hat{\lambda}_1, \ldots, \hat{\lambda}_n$:

$$\text{maximize} \prod_{i=1}^{n} \phi(\hat{\lambda}_i, Y_i)$$
$$\text{subject to} \tag{4}$$
$$|\hat{\lambda}_i - \hat{\lambda}_j| \leq K||X_i - X_j|| \qquad i, j = 1, 2, \ldots, n.$$

Step 2. To get an estimator $\hat{\lambda}$ of the function λ, repeat Step 2 of the regression algorithm, that is, extrapolate the values $\hat{\lambda}_1, \ldots, \hat{\lambda}_n$ at $X_1, \ldots, X_n$ to obtain a Lipschitz continuous function $\hat{\lambda}$ with constant K. Then, given a random observable input X, the estimated probability density function of Y given X is $\phi(\hat{\lambda}(X), y)$.

Remarks:

1. This algorithm tries to maximize the likelihood function, in which instead of a single parameter λ, there is a set of parameters $\lambda_1, \ldots, \lambda_n$ which continuously depend on the input variable X. Namely, this approach finds the maximum likelihood sequence of parameters within the class of parameter sequences satisfying the Lipschitz continuity condition with constant K.

2. Whether the nonlinear programming problem (4) can be solved efficiently depends on the structure of the density function ϕ.

As before, one interesting special case is a time varying statistical model, where the variables $X_1, \ldots, X_n$ stand for the times the outputs $Y_1, \ldots, Y_n$ were observed.

4 Examples

In this section, we apply our DMLE algorithm in several concrete examples. We show how Step 1 of the DMLE algorithm can be performed for these examples. We do not discuss Step 2 in this section since it is the same for all examples.

4.1 Gaussian random variables with unknown mean and constant standard deviation

Suppose that the random values $Y_1, \ldots, Y_n$ are normally distributed with a constant standard deviation σ and *unknown* sequence of means $\mu(X_1), \ldots, \mu(X_n)$. We assume that the function $\mu(x)$ is Lipschitz continuous with *unknown* constant K_0. Using the maximum likelihood approach (4) above, we estimate the function μ by guessing some constant K and solving the following optimization problem in the variables $\hat{\mu}_1, \ldots, \hat{\mu}_n$:

$$\text{maximize} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(Y_i - \hat{\mu}_i)^2}{2\sigma^2} \right)$$
$$\text{subject to}$$
$$|\hat{\mu}_i - \hat{\mu}_j| \leq K||X_i - X_j||, \qquad i, j = 1, 2, \ldots, n.$$

By taking the logarithm of the likelihood function, the problem is equivalent to

$$\text{minimize} \sum_{i=1}^{n} (Y_i - \hat{\mu}_i)^2$$
$$\text{subject to}$$
$$|\hat{\mu}_i - \hat{\mu}_j| \leq K||X_i - X_j||, \qquad i, j = 1, 2, \ldots, n.$$

We recognize this problem as the one described in the previous section for nonlinear regression. We may draw the following analogy with the classical statistical result - given the linear regression model $Y = bX + \epsilon$ with unknown b and a sequence of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ the Least-Squares estimate $\hat{b}$ is also a maximum likelihood estimate, if Y conditioned on X is normally distributed with finite variance.

4.2 Gaussian random variables with unknown mean and unknown standard deviation

Consider a sequence of normally distributed random variables $Y_1, \ldots, Y_n$ with *unknown* means $\mu_1 \equiv \mu(X_1), \ldots, \mu_n \equiv \mu(X_n)$ and *unknown* standard deviations $\sigma_1 \equiv \sigma(X_1), \ldots, \sigma_n \equiv \sigma(X_n)$. We assume that $\mu(x)$ and $\sigma(x)$ are Lipschitz continuous with *unknown* constants K_0^1, K_0^2. Using the maximum likelihood approach (4), we estimate the mean function μ and the standard deviation function σ by guessing constants K_1, K_2 and by solving the following optimization problem in the variables $\hat{\mu}_1, \ldots, \hat{\mu}_n, \hat{\sigma}_1, \ldots, \hat{\sigma}_n$:

$$\text{maximize} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\hat{\sigma}_i} \exp\Big(-\frac{(Y_i - \hat{\mu}_i)^2}{2\hat{\sigma}_i^2}\Big)$$
subject to

$$|\hat{\mu}_i - \hat{\mu}_j| \leq K_1||X_i - X_j||, \qquad i,j = 1,2,\ldots,n,$$
$$|\hat{\sigma}_i - \hat{\sigma}_j| \leq K_2||X_i - X_j||, \qquad i,j = 1,2,\ldots,n.$$

By taking the logarithm of the likelihood function, the above nonlinear programming problem is equivalent to

$$\text{minimize} \sum_{i=1}^{n} \log(\hat{\sigma}_i) + \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{2\hat{\sigma}_i^2}$$
subject to
$$|\hat{\mu}_i - \hat{\mu}_j| \leq K_1||X_i - X_j||, \qquad i,j = 1,2,\ldots,n,$$
$$|\hat{\sigma}_i - \hat{\sigma}_j| \leq K_2||X_i - X_j||, \qquad i,j = 1,2,\ldots,n.$$

Note that here the objective function is not convex.

4.3 Bernoulli random variables

Suppose we observe a sequence of $0,1$ random variables $Y_1,\ldots,Y_n$. Assume that $p(X_i) \equiv \Pr(Y_i = 1)$ depends continuously on some observable variable X_i. In particular, the function $p : \mathcal{X} :\longrightarrow [0,1]$ is Lipschitz continuous, with *unknown* constant K_0. Using the maximum likelihood approach (4) we may construct an approximate function $\hat{p}$ based on observations $(X_1, Y_1),\ldots,(X_n, Y_n)$ by solving the following optimization problem in the variables $\hat{p}_1,\ldots,\hat{p}_n$

$$\text{maximize} \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}$$
subject to

$$|\hat{p}_i - \hat{p}_j| \leq K||X_i - X_j||,\ i,j = 1,2,\ldots,n.$$

By taking the logarithm, this nonlinear programming problem is equivalent to

$$\text{maximize} \sum_{i=1}^{n} Y_i \log(\hat{p}_i) + \sum_{i=1}^{n} (1-Y_i)\log(1-\hat{p}_i)$$
subject to

$$|\hat{p}_i - \hat{p}_j| \leq K||X_i - X_j||,\ i,j = 1,2,\ldots,n.$$

Note that the objective function is concave, and therefore the above nonlinear programming problem is efficiently solvable.

4.4 Exponentially distributed random variables

Suppose we observe a sequence of random values $Y_1,\ldots,Y_n$. Y_i is assumed to be exponentially distributed with rate $\lambda_i = \lambda(X_i)$, and $\lambda(X)$ is a Lipschitz continuous function of the observed input variable X, with *unknown* Lipschitz constant K_0. Using the maximum likelihood approach (4) we may construct an approximate function $\hat{\lambda}$ based on observations $(X_1, Y_1),\ldots,(X_n, Y_n)$ by solving the following optimization problem in the variables $\hat{\lambda}_1,\ldots,\hat{\lambda}_n$:

$$\text{maximize} \prod_{i=1}^{n} \hat{\lambda}_i \exp(-\hat{\lambda}_i Y_i)$$
subject to

$$|\hat{\lambda}_i - \hat{\lambda}_j| \leq K||X_i - X_j||, \qquad i,j = 1,2,\ldots,n.$$

Again by taking the logarithm, this is equivalent to

$$\text{maximize} \sum_{i=1}^{n} \log \hat{\lambda}_i - \sum_{i=1}^{n} \lambda_i Y_i$$
$$\text{subject to} \tag{5}$$

$$|\hat{\lambda}_i - \hat{\lambda}_j| \leq K||X_i - X_j||, \qquad i,j = 1,2,\ldots,n.$$

This nonlinear programming problem is also efficiently solvable, since the objective is a concave function.

5 Simulation results

In this section, we provide some simulation results involving the Regression algorithm from Section 2. We also compare the performances of the Regression algorithm and kernel regression on the same samples of artificially generated data. The resulting approximating function $\hat{f}^n$ is measured against the underlying regression function f.

Let us consider a particular case of the model from Section 2,

$$Y = \sin(X) + \psi$$

with $0 \leq X \leq 2\pi$ and noise term ψ normally distributed as $N(0, \sigma^2)$. We divide the interval $[0, 2\pi]$ into $n-1$ equal intervals and pick end points of these intervals $X_i = 2\pi(i-1)/(n-1), i = 1,\ldots,n$. We generate n independent noise terms $\psi_1, \psi_2,\ldots,\psi_n$ with normal $N(0,\sigma^2)$ distribution. After running Step 1 of the Regression Algorithm on the values $X_i, Y_i = \sin(X_i) + \psi_i, i = 1,2,\ldots,n$ we obtain approximating values $\hat{f}_1,\ldots,\hat{f}^n$. We also compute kernel regression estimates of the function $\sin(x), x \in [0, 2\pi]$ using the same samples $(X_i, Y_i),\ i = 1,2,\ldots,n$. For the estimating functions $\hat{f}$ obtained by either the Regression algorithm or kernel regression, we use performance measures

$$d_1 \equiv \max_{1\leq i\leq n} |\hat{f}(X_i) - \sin(X_i)|$$

and

$$d_2 \equiv \Big(\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(X_i) - \sin(X_i))^2\Big)^{\frac{1}{2}}$$

The first performance measure approximates the uniform (maximal) distance $\max_{0 \le x \le 2\pi} |\hat{f}(x) - \sin(x)|$ between the regression function $\sin(x)$ and its estimate $\hat{f}$. In Section 6 we will present some theoretical results on the distribution of the distance $\max_{0 \le x \le 2\pi} |\hat{f}(x) - f(x)|$ for any Lipschitz continuous function $f(x)$. The second performance measure approximates the average distance between $\sin(x)$ and $\hat{f}(x)$.

We summarize the results of these experiments in Tables 1 and 2, corresponding to sample sizes $n = 50$ and $n = 100$ and performance measure d_1, and in Table 3, corresponding to sample size $n = 100$ and the performance measure d_2. Each row corresponds to a different standard deviation σ used for the experiment. The first two columns list the values of the performance d obtained by the Regression algorithm using Lipschitz constants $K = 1$ and $K = 2$. Note, that the function $\sin(x)$ has Lipschitz constant $K_0 = 1$. That is, $K_0 = 1$ is the smallest value K_0, for which $|\sin(x) - \sin(y)| \le K_0|x - y|$ for all $x, y \in [0, 2\pi]$. The last two columns are the results of kernel regression estimation using the same data samples and bandwidths $\delta = 0.3$ and $\delta = 0.1$. We use $\phi(x, x_0) = e^{-\frac{(x-x_0)^2}{\delta^2}}$ as a kernel function.

$n = 50$	Regression algorithm		Kernel regression	
σ	$K = 1$	$K = 2$	$\delta = .3$	$\delta = .1$
0.5	0.5775	0.6252	0.4748	0.8478
0.1	0.2144	0.2114	0.2002	0.1321
0.05	0.1082	0.1077	0.1523	0.1140
0.01	0.0367	0.0211	0.1346	0.0503
0.001	0.0026	0.0027	0.1284	0.0453

Table 1. Performance measure d_1

$n = 100$	Regression algorithm		Kernel regression	
σ	$K = 1$	$K = 2$	$\delta = .3$	$\delta = .1$
0.5	0.2861	0.2617	0.2340	0.4762
0.1	0.1100	0.1438	0.1566	0.1061
0.05	0.0766	0.0810	0.1411	0.0773
0.01	0.0200	0.0273	0.1525	0.0682
0.001	0.0026	0.0025	0.1475	0.0618

Table 2. Performance measure d_1

$n = 100$	Regression algorithm		Kernel regression	
σ	$K = 1$	$K = 2$	$\delta = .3$	$\delta = .1$
0.5	0.1299	0.2105	0.1157	0.1868
0.1	0.0515	0.0688	0.0618	0.0569
0.05	0.0272	0.0433	0.0574	0.0519
0.01	0.0093	0.0101	0.0575	0.0575
0.001	0.0008	0.0010	0.0566	0.0567

Table 3. Performance measure d_2

Examining the performance of the Regression algorithm for the choices $K = 1$ and $K = 2$, we see that the algorithm is not particularly sensitive to the deviation of the chosen constant K from the correct constant K_0. The values obtained with $K = 1$ and $K = 2$ are quite close to each other. Metric d_1 is a more conservative measure of accuracy than a metric d_2. Therefore, it is not surprizing that the approximating errors in Table 2 are bigger then the corresponding errors in Table 3.

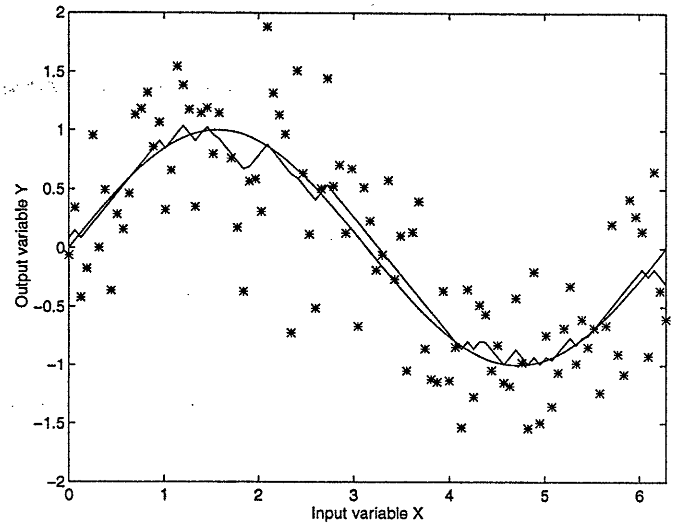

Figure 1: The Regression algorithm

Also, it seems that for each choice of the bandwidth δ there are certain values of σ for which the performance of the two algorithms is the same, or the performance of kernel regression is slightly better ($\sigma = 0.5, 0.1$ for $\delta = 0.3$; $\sigma = 0.1, 0.05$ for $\delta = 0.1$). However, as the noise level σ becomes smaller, we see that the Regression algorithm outperforms kernel regression. This is consistent with Remark 6 in Section 2: the Regression algorithm is more robust with respect to small noise levels.

In figure 1 we have plotted the results of running the Regression algorithm on a data sample, generated using the model above. The sample size used is $n = 100$, and the standard deviation of the noise is $\sigma = .5$. The piecewise linear curve around the curve $\sin(x)$ is the resulting approximating function $\hat{f}$. The "*"-s are the actual observations (X_i, Y_i), $i = 1, 2, \ldots, 100$. We see that the algorithm is successful in obtaining a fairly close approximation of the function $\sin(x)$.

6 Convergence to the true regression function; consistency result.

In this section, we discuss the consistency of our convex optimization regression algorithm for the case of one dimensional input and output variables X, Y. Roughly speaking, we show that for the nonlinear regression model $Y = f(X) + \psi$ in Section 1, the estimated function $\hat{f}$ constructed by the regression algorithm, converges to the true function f as the number of observations goes to infinity, if X and Y are bounded random variables and our constant K is *bigger* than the true constant K_0. For any continuous function g defined on a closed interval $I \subset \mathcal{R}$, let the norm $||g||_\infty$ be defined as $\max_{x \in I} |g(x)|$.

Theorem 6.1 *Consider bounded random variables $X, Y \in \Re$, $a_1 \le X \le a_2$, $b_1 \le Y \le b_2$, described by joint probability distribution function $F(x,y)$. Suppose that $f(x) \equiv E[Y|X = x]$ is a Lipschitz continuous function, with constant K_0, and suppose that the distribution of the random variable X has a continuous positive density function.*

For any sample of i.i.d. outcomes $(X_1, Y_1), \ldots, (X_n, Y_n)$, and a constant $K > 0$, let $\hat{f}^n \equiv \hat{f}$ be the estimating function computed by the regression algorithm of Section 2.

If $K \ge K_0$, then

1. *$\hat{f}^n$ converges to f uniformly and almost surely. That is,*

$$\lim_{n \to \infty} ||\hat{f}^n - f||_\infty = 0, \qquad \text{w.p.1.}$$

2. *For any $\epsilon > 0$, there exist constants $\gamma_1 = \gamma_1(\epsilon)$ and $\gamma_2 = \gamma_2(\epsilon)$ such that*

$$\Pr\Big\{||\hat{f}^n - f||_\infty > \delta\Big\} \le \gamma_1 e^{-\gamma_2 n}, \qquad \text{for all } n.$$

Proof: Let $\Im$ be the set of all Lipschitz continuous functions $\hat{f} : [a_1, a_2] \to [b_1, b_2]$ with constant K. Introduce the risk function

$$Q(x, y, \hat{f}) = (y - \hat{f}(x))^2$$

for every $(x, y) \in [a_1, a_2] \times [b_1, b_2]$, $\hat{f} \in \Im$. Then the solution $\hat{f}^n$ obtained from steps 1 and 2 of the Regression algorithm is a solution to the problem

$$\text{Minimize}_{\hat{f} \in \Im} \frac{1}{n} \sum_{i=1}^{n} Q(X_i, Y_i, \hat{f}) \tag{6}$$

- the Empirical Risk Minimization problem (see [1],p.18). Notice also that the regression function f is a solution to the minimization problem

$$\text{Minimize}_{\hat{f} \in \Im} \int Q(x, y, \hat{f}) dF(x, y)$$

since for each fixed $x \in [a_1, a_2]$ the minimum of $E[(Y - \hat{f}(x))^2 | X = x]$ is achieved by $\hat{f}(x) = f(x)$.

Much of our proof of the Theorem 6.1 is built on the concept of *VC entropy* introduced first by Vapnik and Chervonenkis. For any fixed sequence

$$(x_1, y_1), \ldots, (x_n, y_n) \in [a_1, a_2] \times [b_1, b_2]$$

consider the set of vectors

$$\{(Q(x_1, y_1, \hat{f}), \ldots, Q(x_n, y_n, \hat{f})),\ \hat{f} \in \Im\} \tag{7}$$

obtained by varying $\hat{f}$ over $\Im$.
Let $N(\epsilon, \Im, (x_1, y_1), \ldots, (x_n, y_n))$ be the number of elements (the cardinality) of a minimal ϵ-net of this set of vectors. That is $N(\epsilon, \Im, (x_1, y_1), \ldots, (x_n, y_n))$ is the smallest integer k, for which there exist k vectors

$$q_1, q_2, \ldots, q_k \in \mathcal{R}^n,$$

such that for any vector q in the set (7), $||q - q_j||_\infty < \epsilon$ for some $j = 1, 2, \ldots, k$, where $||\cdot||_\infty$ is the maximum norm in $\mathcal{R}^n$. The following definition of VC entropy was used by Haussler in [6].

Definition 6.1 *For any $\epsilon > 0$, the VC entropy of $\Im$ for samples of size n is defined to be*

$$H^{\Im}(\epsilon, n) \equiv E[N(\epsilon, \Im, (X_1, Y_1), \ldots, (X_n, Y_n))]$$

The following theorem is a variation of Pollard's result (Theorem 24, page 25, [5]) and was proven by Haussler (Corollary 1, page 11, [6]).

Proposition 6.1 *There holds*

$$\Pr\Big\{\sup_{\hat{f} \in \Im} \Big| \int Q(x, y, \hat{f}) dF(x, y) - \frac{1}{n} \sum_{i=1}^{n} Q(X_i, Y_i, \hat{f}) \Big| > \epsilon\Big\}$$
$$\le 4 H^{\Im}(\epsilon, n) e^{-\epsilon^2 n / 64 (b_2 - b_1)^4}$$

The key to our analysis is to show that for the case of class $\Im$ of Lipschitz continuous functions with Lipschitz constant K, the right-hand side of the inequality above converges to zero as the sample size n goes to infinity. The following proposition achieves this goal by showing that the VC entropy of $\Im$ is finite, independently of the sample size n.

Proposition 6.2 *For each $\epsilon > 0$ and sequence $(x_1, y_1), \ldots, (x_n, y_n)$ from $[a_1, a_2] \times [b_1, b_2]$ there holds*

$$N(\epsilon, \Im, (x_1, y_1), \ldots, (x_n, y_n))$$
$$\le \Big(\frac{6(b_2 - b_1)^2}{\epsilon} + 1\Big) 3^{\frac{6K(a_2 - a_1)(b_2 - b_1)}{\epsilon} + 1}$$

Proof: see the Appendix.

Combining Propositions 6.2 and 6.1, we conclude

Proposition 6.3 *There holds*

$$\Pr\Big\{\sup_{\hat{f} \in \Im} \Big| \int Q(x, y, \hat{f}) dF(x, y) - \frac{1}{n} \sum_{i=1}^{n} Q(X_i, Y_i, \hat{f}) \Big| > \epsilon\Big\}$$

$$\leq 4\Big(\frac{6(b_2-b_1)^2}{\epsilon}+1\Big)3^{\frac{6K(a_2-a_1)(b_2-b_1)}{\epsilon}+1}e^{-\epsilon^2 n/64(b_2-b_1)^4}.$$

In particular,

$$\Pr\Big\{\sup_{\hat{f}\in\Im}\Big|\int Q(x,y,\hat{f})dF(x,y)-\frac{1}{n}\sum_{i=1}^{n}Q(X_i,Y_i,\hat{f})\Big|>\epsilon\Big\}$$

$$\longrightarrow 0, \qquad \text{as } n\to\infty.$$

We have proved that the difference between the risk $\int Q(x,y,\hat{f})dF(x,y)$ and the empirical risk $\frac{1}{n}\sum_{i=1}^{n}Q(X_i,Y_i,\hat{f})$ converges to zero uniformly in probability for our class $\Im$. Let the norm $||\cdot||_2$ be defined for any function $g\in\Im$ as

$$||g||_2=\Big(\int g^2dF(X)\Big)^{1/2}$$

We now use Proposition 6.3 to prove that the tail of the distribution of the difference $||\hat{f}^n-f||_2$ converges to zero exponentially fast.

Proposition 6.4 *There holds*

$$\Pr\Big\{||\hat{f}^n-f||_2>\epsilon\Big\} \tag{8}$$

$$\leq 8\Big(\frac{12(b_2-b_1)^2}{\epsilon^2}+1\Big)3^{\frac{12K(a_2-a_1)(b_2-b_1)}{\epsilon^2}+1}e^{-\frac{\epsilon^4}{256(b_2-b_1)^4}n}$$

Proof: see the Appendix.

Our last step is to show that $||\hat{f}^n-f||_\infty\longrightarrow 0$ almost surely. For any $\epsilon>0$ introduce

$$\alpha(\epsilon)\equiv\inf_{a_1\leq x_0\leq a_2}\Pr\{|X-x_0|<\epsilon\} \tag{9}$$

The next lemma is an immediate consequence of an assumption that the distribution of X is described by a positive and continuous density function.

Lemma 6.1 *For every $\epsilon>0$ there holds $\alpha(\epsilon)>0$.*

Proof: The function $\alpha(x_0,\epsilon)\equiv\Pr\{|X-x_0|<\epsilon\}$ is continuous and positive, since, by assumption, the distribution of X has a positive density function. Since this function is defined on a compact set $[a_1,a_2]$ it assumes a positive minimum $\alpha(\epsilon)$. □

The following lemma proves that convergence in $||\cdot||_2$ norm implies the convergence in $||\cdot||_\infty$ for the class $\Im$ of Lipschitz continuous functions with constant K. It will allow us to prove the result similar to (8) but for the distance $||\hat{f}^n-f||_\infty$.

Lemma 6.2 *Consider a Lipschitz continuous function g on $[a_1,a_2]$ with Lipschitz constant K. Suppose that for some $\epsilon>0$ there holds $||g||_\infty\geq\epsilon$. Then $||g||_2\geq\frac{1}{2}\epsilon\alpha^{\frac{1}{2}}(\epsilon/2K)>0$. In particular, for a sequence $g,g_1,\ldots,g_n,\ldots$ of Lipschitz continuous functions with a common Lipschitz constant K, $||g_n-g||_2\to 0$ implies $||g_n-g||_\infty\to 0$.*

Proof: Suppose $||g||_\infty\geq\epsilon$. That is, for some $a\in[a_1,a_2]$, $|g(a)|\geq\epsilon$. Set $\delta=\epsilon/(2K)$. We have

$$||g||_2^2\geq\int_{(a-\delta,a+\delta)}g^2(x)dF(x)$$

For any $x\in(a-\delta,a+\delta)$ we have $|g(x)-g(a)|\leq K\delta$. It follows that $|g(x)|\geq\epsilon-K\delta=\epsilon/2$, for all $x\in(a-\delta,a+\delta)$. Therefore,

$$||g||_2^2\geq(\epsilon/2)^2\Pr\Big\{a-\epsilon/(2K)\leq X\leq a+\epsilon/(2K)\Big\}$$

$$\geq\frac{\epsilon^2}{4}\alpha(\epsilon/2K)>0$$

where the last inequality follows from Lemma 6.1. □

We use Lemma 6.2 to prove our final proposition:

Proposition 6.5 *There holds*

$$\Pr\Big\{||\hat{f}^n-f||_\infty>\epsilon\Big\}\leq \tag{10}$$

$$\leq 8\Big(\frac{48(b_2-b_1)^2}{\epsilon^2\alpha(\epsilon/2K)}+1\Big)3^{\frac{48K(a_2-a_1)(b_2-b_1)}{\epsilon^2\alpha(\epsilon/2K)}+1}e^{-\frac{\epsilon^4\alpha^2(\epsilon/2K)}{2^{12}(b_2-b_1)^4}n}$$

where $\alpha(\epsilon)$ is defined by (9).

Proof: Note from Lemma 6.2

$$\Pr\Big\{||\hat{f}^n-f||_\infty>\epsilon\Big\}\leq\Pr\Big\{||\hat{f}^n-f||_2>\frac{1}{2}\epsilon\alpha^{\frac{1}{2}}(\epsilon/2K)\Big\}$$

Then the result follows immediately from the Proposition 6.4 □

Proposition 6.5 establishes the convergence $||\hat{f}^n-f||_\infty\to 0$ in probability. We now set

$$\gamma_1=8\Big(\frac{48(b_2-b_1)^2}{\epsilon^2\alpha(\epsilon/2K)}+1\Big)3^{\frac{48K(a_2-a_1)(b_2-b_1)}{\epsilon^2\alpha(\epsilon/2K)}+1}$$

and

$$\gamma_2=\frac{\epsilon^4\alpha^2(\epsilon/2K)}{2^{12}(b_2-b_1)^4}$$

To complete the proof of the theorem, we need to establish almost sure convergence of $\hat{f}^n$ to f. But this is a simple consequence of the exponential bound (10) and the Borel-Cantelli lemma.

Theorem 6.1 is proved. □

The bound (10) provides us with a confidence interval on the estimate $\hat{f}^n$. Given the training sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, we construct the estimate $\hat{f}^n = \hat{f}^n(X_1, Y_1, \ldots, X_n, Y_n)$. Then given an arbitrary input observation $X \in [a_1, a_2]$ the probability that the deviation of the estimated output $\hat{f}^n(X)$ from the true output $f(X)$ is more than ϵ, is smaller than the right-hand side of the inequality (10) Note, that the bound (10) depends only on the distribution of X and not on the distribution of $Y|X$. Unfortunately the constants γ_1 and γ_2 are too large for practical purposes. Our simulation results from the Section 5 suggest that the rate of convergence $\hat{f}^n \to f$ is better than the very pessimistic ones in Propositions 6.4 and 6.5. It would be interesting to investigate whether better rates of convergence can be established, with corresponding upper bounds more practically useful.

7 Conclusions

We have proposed a convex optimization approach to the non-parametric regression estimation problem. A number of desirable properties were proven for this technique: average unbiasedness, and a strong form of consistency.

We have also proposed an optimization approach for the maximum likelihood estimation of dynamically changing parameters in statistical models. For many classical distributional forms, the objective function in the optimization problem is convex, and the constraints are linear. These problems are therefore efficiently solvable. It would be interesting to investigate any consistency property of this estimation procedure. The other question for further investigation seems to be the selection of the constant K. A good choice of the constant K is crucial for the approximation to be practically successful. Finally, it is of interest to improve the rates of convergence provided by the Proposition 6.5

8 Appendix

We provide in this appendix the proofs of the most technical parts of the paper.

Proposition 6.2 *Proof:* Fix $\epsilon > 0$ and $x_1, y_1, \ldots, x_n, y_n$. Let

$$\delta = \delta(\epsilon) = \frac{\epsilon}{6(b_2 - b_1)}. \tag{11}$$

Divide the interval $[a_1, a_2]$ into $m_1 = \lfloor K(a_2 - a_1)/\delta \rfloor + 1$ equal size intervals. Here $\lfloor \cdot \rfloor$ stands for the largest integer not bigger than x. Clearly the size of each interval is smaller than δ/K. Let $L_1, L_2, \ldots, L_{m_1}$ be the left endpoints of these intervals and let L_{m_1+1} be the right endpoint of the interval I_{m_1}. Divide also interval $[b_1, b_2]$ into $m_2 = \lfloor (b_2 - b_1)/\delta \rfloor + 1$ equal size intervals $J_1, \ldots, J_{m_2}$ All the intervals J_j, $j = 1, 2, \ldots, m_2$ have lengths less than δ. Let $P_1, P_2, \ldots, P_{m_2}$ be the left endpoints of the intervals $J_1, J_2, \ldots, J_{m_2}$.

We prove the proposition by explicitly constructing an ϵ-net of the set (7). We start by building $\delta(\epsilon)$ net of the set $\Im$ with respect to the $||\cdot||_\infty$ norm. Consider the set of all functions $g \in \Im_0$ which satisfy the following three conditions

1. $g(L_i) \in \{P_1, P_2, \ldots, P_{m_1}\}, \qquad i = 1, 2, \ldots, m_1 + 1.$
2. If $g(L_i) = P_j$ then
$$g(L_{i+1}) \in \{P_{j-1}, P_j, P_{j+1}\}, \qquad i = 1, 2, \ldots, m_1.$$
3. For all $x \in (L_i, L_{i+1})$,
$$g(x) = \mu g(L_i) + (1 - \mu) g(L_{i+1}),$$
where
$$\mu = \frac{L_{i+1} - x}{L_{i+1} - L_i}.$$
That is, the function g is obtained by linear extrapolation of the values $g(L_1), g(I_2), \ldots, g(L_{m_1+1})$.

It is not hard to see that the family $\Im_0$ is a non-empty set of Lipschitz continuous functions with constant K. The latter fact is guaranteed by condition 2 and the fact $L_{i+1} - L_i < \delta/K$, $P_{j+1} - P_j < \delta$. The cardinality of this set satisfies

$$|\Im_0| \le m_2 3^{m_1}$$

since we have m_2 choices for $g(I_1)$, and only three choices for each subsequent value $g(I_2), \ldots, g(I_{m_1+1})$ by condition 2.

Consider now an arbitrary function $f \in \Im$. We will construct a function $g_f \in \Im_0$ such that $||g_f - f||_\infty < 3\delta$. For $i = 1, 2, \ldots, m_1 + 1$, set $g_f(L_i) = P_j$ if $f(L_i) \in J_j$, $j = 1, 2, \ldots, m_2$. Then linearly extrapolate the values $g_f(L_i)$ to get the function $g_f : [a_1, a_2] \to [b_1, b_2]$. Clearly g_f satisfies the conditions 1 and 3 of the set $\Im_0$. Also, since f is Lipschitz continuous with the constant K, then $|f(L_{i+1}) - f(L_i)| < \delta$. It follows that $f(L_{i+1}) \in J_{j-1} \cup J_j \cup J_{j+1}$. Therefore $g_f(L_{i+1}) \in \{P_{j-1}, P_j, P_{j+1}\}$. Thus condition 2 is also satisfied. Finally, suppose $x \in [L_i, L_{i+1}]$. Then

$$|f(x) - g_f(x)| \le |f(x) - f(L_i)| + |f(L_i) - g_f(L_i)|$$
$$+ |g_f(L_i) - g(x)| \le \delta + \delta + \delta = 3\delta.$$

We see that $||f - g_f||_\infty < 3\delta$ and, as a result, the set $\Im_0$ is a 3δ-net of the family $\Im$.

We now show that for each fixed sequence $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the set of vectors

$$\begin{aligned} &\{(Q(x_1, y_1, g), \ldots, Q(x_n, y_n, g), \ g \in \Im_0\} \\ &= \{((y_1 - g(x_1))^2, \ldots, (y_n - g(x_n))^2, \ g \in \Im_0\} \end{aligned} \tag{12}$$

is an ϵ-net of the set

$$\{(Q(x_1, y_1, f), \ldots, Q(x_n, y_n, f), \ f \in \Im\} \tag{13}$$

In fact, for each $f \in \Im$ and $g_f \in \Im_0$ satisfying $||f - g_f||_\infty < 3\delta$, and $i = 1, 2, \ldots, n$, we have

$$|(y_i - f(x_i))^2 - (y_i - g_f(x_i))^2|$$
$$= |f(x_i) - g_f(x_i)| \cdot |2y_i - f(x_i) - g_f(x_i)|$$
$$\leq 3\delta 2|b_2 - b_1| = \epsilon$$

and the set (12) is an ϵ-net of the set (13). The cardinality of the set (12) is no bigger than

$$m_2 3^{m_1} = \Big(\lfloor\frac{b_2 - b_1}{\delta(\epsilon)}\rfloor + 1\Big)3^{\lfloor\frac{K(a_2-a_1)}{\delta(\epsilon)}\rfloor+1}$$
$$\leq \Big(\frac{b_2 - b_1}{\delta(\epsilon)} + 1\Big)3^{\frac{K(a_2-a_1)}{\delta(\epsilon)}+1}$$

We have proved

$$N(\epsilon, \Im, (x_1, y_1), \ldots, (x_n, y_n))$$
$$\leq \Big(\frac{b_2 - b_1}{\delta(\epsilon)} + 1\Big)3^{\frac{K(a_2-a_1)}{\delta(\epsilon)}+1}$$

The statement of the Proposition then follows from (11).□

Proposition 6.4 *Proof:* The identity

$$\int (y - \hat{f}(x))^2 dF(x, y)$$
$$= \int (y - f(x))^2 dF(x, y) + \int (f(x) - \hat{f}(x))^2 dF(x, y) \tag{14}$$
$$= \int (y - f(x))^2 dF(x, y) + ||\hat{f} - f||_2^2$$

can be easily established for any $\hat{f} \in \Im$ by using the fact

$$(y - \hat{f}(x))^2 = (y - f(x))^2$$
$$+2(y - f(x))(f - \hat{f}(x)) + (f(x) - \hat{f}(x))^2$$

and the orthogonality property

$$E\Big[(Y - f(X))(f(X) - \hat{f}(X)\Big] = 0$$

We have

$$\Pr\Big\{||\hat{f}^n - f||_2^2 > \epsilon\Big\}$$
$$= \Pr\Big\{||\hat{f}^n - f||_2^2 + \int Q(x, y, f)dF(x, y) - \frac{1}{n}\sum_{i=1}^{n} Q(X_i, Y_i, \hat{f}^n)$$
$$-\Big(\int Q(x, y, f)dF(x, y) - \frac{1}{n}\sum_{i=1}^{n} Q(X_i, Y_i, \hat{f}^n)\Big) > \epsilon\Big\}$$

The following upper bound then holds

$$\Pr\Big\{||\hat{f}^n - f||_2^2 > \epsilon\Big\} \leq \Pr\Big\{||\hat{f}^n - f||_2^2 + \int Q(x, y, f)dF(x, y) -$$
$$-\frac{1}{n}\sum_{i=1}^{n} Q(\hat{f}^n, X_i, Y_i) > \epsilon/2\Big\} \tag{15}$$
$$+\Pr\Big\{\int Q(x, y, f)dF(x, y) - \frac{1}{n}\sum_{i=1}^{n} Q(\hat{f}^n, X_i, Y_i) < -\epsilon/2\Big\}$$

From the decomposition (14) we have

$$\int Q(x, y, \hat{f}^n)dF(x, y) = \int Q(x, y, f)dF(x, y) + ||\hat{f}^n - f||_2^2$$

Therefore

$$\Pr\Big\{||\hat{f}^n - f||_2^2 + \int Q(x, y, f)dF(x, y)$$
$$-\frac{1}{n}\sum_{i=1}^{n} Q(\hat{f}^n, X_i, Y_i) > \epsilon/2\Big\}$$
$$= \Pr\Big\{\int Q(\hat{f}^n, x, y)dF(x, y) - \frac{1}{n}\sum_{i=1}^{n} Q(\hat{f}^n, X_i, Y_i) > \epsilon/2\Big\}$$

From Proposition 6.3

$$\Pr\Big\{\int Q(\hat{f}^n, x, y)dF(x, y) - \frac{1}{n}\sum_{i=1}^{n} Q(\hat{f}^n, X_i, Y_i) > \epsilon/2\Big\} \tag{16}$$
$$\leq 4\Big(\frac{12(b_2 - b_1)^2}{\epsilon} + 1\Big)3^{\frac{12K(a_2-a_1)(b_2-b_1)}{\epsilon}+1}e^{-\epsilon^2 n/256(b_2-b_1)^4}$$

Also from (6)

$$\sum_{i=1}^{n} Q(\hat{f}^n, X_i, Y_i) \leq \sum_{i=1}^{n} Q(f, X_i, Y_i)$$

As a result

$$\Pr\Big\{\int Q(x, y, f)dF(x, y) - \frac{1}{n}\sum_{i=1}^{n} Q(\hat{f}^n, X_i, Y_i) < -\epsilon/2\Big\}$$
$$\leq \Pr\Big\{\int Q(x, y, f)dF(x, y) - \frac{1}{n}\sum_{i=1}^{n} Q(f, X_i, Y_i) < -\epsilon/2\Big\}$$

Again from Proposition 6.3,

$$\Pr\Big\{\int Q(f, x, y)dF(x, y) - \frac{1}{n}\sum_{i=1}^{n} Q(f, X_i, Y_i) < -\epsilon/2\Big\} \tag{17}$$
$$\leq 4\Big(\frac{12(b_2 - b_1)^2}{\epsilon} + 1\Big)3^{\frac{12K(a_2-a_1)(b_2-b_1)}{\epsilon}+1}e^{-\epsilon^2 n/256(b_2-b_1)^4}$$

It follows from (15),(16), and (17)

$$\Pr\Big\{||\hat{f}^n - f||_2^2 > \epsilon\Big\} \leq$$
$$8\Big(\frac{12(b_2 - b_1)^2}{\epsilon} + 1\Big)3^{\frac{12K(a_2-a_1)(b_2-b_1)}{\epsilon}+1}e^{-\epsilon^2 n/256(b_2-b_1)^4}$$

This completes the proof.□

References

[1] V. Vapnik. Nature of Learning Theory. Springer-Verlag, New York, 1996.

[2] R. Eubank. Spline Smoothing and Nonparametric Regression. M.Dekker, New York, 1988.

[3] L. P. Devroye. The uniform convergence of nearest neighbor regression function estimators and their application in optimization. *IEEE Trans.Inform.Theory,* 24, 142-151, 1978.

[4] M. Bazaara, H. Sherali, C. Shetti. Nonlinear Programming; Theory and Algorithms. Wiley, New York, 1993.

[5] Pollard. Convergence of Stochastic Processes. Springer-Verlag, 1984.

[6] D.Haussler. Generalizing the PAC Model for Neural Net and Other Learning Applications. University of California Santa Cruz Technical Report UCSC-CRL-89-30.