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

Dimitris Bertsimas * David Gamarnik † John N. Tsitsiklis ‡

January 26, 1999

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 the minimization of the sum of empirical squared errors, subject to the constraints implied by Lipschitz continuity. The resulting optimization problem has a convex objective function and linear constraints, and as a result, is efficiently solvable. The estimated 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 the 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 the function f based on a sequence of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. 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 observe $Y_1, Y_2, \ldots, Y_n$, and the problem is to compute the 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.

*Sloan School of Management and Operations Research Center, MIT Cambridge, MA 02139. dbertsim@aris.mit.edu. Research partially supported by a Presidential Young Investigator Award DDM-9158118, with matching funds from Draper Laboratory.

†Operations Research Center, MIT Cambridge, MA 02139. Research partially supported by the ARO under grant DAAL-03-92-G-0115 and by the NSF under grant DDM-9158118. Current affiliation: IBM T.J.Watson Research Center, P.O.Box 218, Yorktown Heights, NY 10598. gamarnik@watson.ibm.com.

‡Laboratory for Information and Decision Systems and Operations Research Center, MIT Cambridge, MA 02139. jnt@mit.edu. Research partially supported by the ARO under grant DAAL-03-92-G-0115.

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 the 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 $O(\frac{1}{n^{2/3}})$ 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 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 estimated function $\hat{f}^n \equiv \hat{f}(\cdot; X_1, Y_1, \ldots, X_n, Y_n)$ 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 simulation results in Section 5 and Theorem 2 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 estimated 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, in principle, to build confidence intervals around $\hat{f}^n$. However, the constants in our estimates of the tail probabilities are too large to be practically useful.

2 A nonlinear regression model

In this section, we demonstrate how convex optimization algorithms can be used for nonlinear regression analysis. Let X be a random vector taking values in a set $\mathcal{X} \subset \Re^m$ and let Y be a random variable taking values in a set $\mathcal{Y} \subset \Re$. We are given a model (1) in which the function $f : \mathcal{X} \mapsto \mathcal{Y}$ is Lipschitz continuous with some unknown parameter K. Namely, $|f(x_1) - f(x_2)| \le K\|x_1 - x_2\|_\infty$ for all $x_1, x_2 \in \mathcal{X}$. Throughout the paper, $\|\cdot\|_\infty$ is used to denote the maximum norm on $\Re^m$. That is, $||x||_\infty = \max_i |x_i|$, for all $x \in \Re^d$. The objective is to find an estimate $\hat{f}$ of the true function f based on the sequence of noisy observations. We consider a model of the form $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$:

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

We assume that the random variables $\psi_1, \ldots, \psi_n$, conditioned on $X_1, \ldots, X_n$, have zero mean and are mutually independent. 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 $\hat{f}_1, \ldots, \hat{f}_n$:

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

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 $\hat{f}_1, \ldots, \hat{f}_n$ obtained in Step 1, to a Lipschitz continuous function $\hat{f} : \mathcal{X} \mapsto \Re$ with the constant K, as follows: for any $x \in \mathcal{X}$, let

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

The following proposition justifies Step 2 of the above algorithm.

Proposition 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 $i = \text{argmax}_{1 \le j \le n}\{\hat{f}_j - K||x_1 - X_j||_\infty\}$, i.e., $\hat{f}(x_1) = \hat{f}_i - K||x_1 - X_i||_\infty$. Moreover, by the definition of $\hat{f}(x_2)$, $\hat{f}(x_2) \ge \hat{f}_i - K||x_2 - X_i||_\infty$. Therefore,

$$\begin{aligned} \hat{f}(x_1) - \hat{f}(x_2) \quad & \le \quad \hat{f}_i - K||x_1 - X_i||_\infty - (\hat{f}_i - K||x_2 - X_i||)_\infty \\ & = \quad K||x_2 - X_i||_\infty - K||x_1 - X_i||_\infty \\ & \le K||x_2 - x_1||_\infty. \end{aligned}$$

By a symmetric argument, we obtain

$$\hat{f}(x_2) - \hat{f}(x_1) \le K||x_2 - x_1||_\infty.$$

For $x = X_i$, we have $\hat{f}_i - K||x - X_i||_\infty = \hat{f}_i$. For all $j \ne i$, constraint (2) guarantees that $\hat{f}_j - K||x - X_j||_\infty \le \hat{f}_i$. It follows that $\hat{f}(X_i) = \hat{f}_i$. $\square$

In Step 2, we could take instead

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

or

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

Proposition 1 holds for 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 $\hat{f}_1, \ldots, \hat{f}_n$:

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{n} (Y_i - \hat{f}_i)^2 \\ &\text{subject to} && |\hat{f}_{i+1} - \hat{f}_i| \le K(t_{i+1} - t_i), \qquad i = 1, 2, \ldots, n-1. && (3) \\ & && && (4) \end{aligned}$$

Step 2'. The extrapolation step can be performed in the following way. For any t, with $t_i \le t < t_{i+1}$, 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 proposed algorithm relies on the minimization of the sum of the empirical squared errors between the estimated function value $\hat{f}_i$ at point X_i and the observation Y_i, in such a way that the estimates $\hat{f}_1, \ldots, f_n$ 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 $K \ge K_0$, 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., 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. The optimization problems (2) or (3) are quadratic programming problems, involving a convex quadratic objective function and linear constraints, and can be efficiently solved (see [4]). In fact, interior point methods can find optimal solutions in polynomial time.

5. Setting $K = 0$, yields the 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 zero, then the estimated function $\hat{f}$ coincides with the true function f at the observed input values:

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

   This compares favorably with the kernel regression and smoothing spline techniques, where due to the selected positive bandwidth or positive regularization parameter respectively, the estimated function is not equal to the true function even if the noise is zero. Thus, our method is more 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 1 *Let the estimates $\hat{f}_i$ be obtained from the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, according to 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 estimates $\hat{f}_1, \ldots, \hat{f}_n$ be obtained using Step 1 of the regression algorithm. Observe that the estimates $\hat{f}_i + c,\ i = 1, 2, \ldots, n$, also satisfy the constraints in (2), for any $c \in \Re$. Since the first set of estimates is optimal, we must have

$$\sum_{i=1}^{n}(Y_i - \hat{f}_i)^2 \le \sum_{i=1}^{n}(Y_i - \hat{f}_i - c)^2, \qquad \forall\ c \in \Re.$$

Taking the derivative of the right-hand side with respect to c, and setting it to zero at $c = 0$, we obtain

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

or

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

It follows that

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

where the last step follows from the zero mean property of the random variables ψ_i. $\square$

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.

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

More precisely, conditioned on X_i, the random variable 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 observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. As a solution we propose the following algorithm:

Dynamic Maximum Likelihood Estimation (DMLE) Algorithm:

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

$$\begin{aligned} &\text{maximize} && \prod_{i=1}^{n} \phi(\hat{\lambda}_i, Y_i) \\ &\text{subject to} && |\hat{\lambda}_i - \hat{\lambda}_j| \leq K||X_i - X_j||_\infty, \qquad i, j = 1, 2, \ldots, n. \end{aligned} \tag{5}$$

Step 2. To get an estimate $\hat{\lambda}$ 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 depend continuously 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 (5) can be solved efficiently or not 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 at which the outputs $Y_1, \ldots, Y_n$ were observed.

4 Examples

In this section, we apply our DMLE algorithm to several concrete examples and show how Step 1 can be carried out. We do not discuss Step 2 in this section since it is always the same.

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 (5), we estimate the function μ by guessing some constant K and solving the following optimization problem in the variables $\hat{\mu}_1, \ldots, \hat{\mu}_n$:

$$\begin{array}{ll} \text{maximize} & \displaystyle\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(Y_i - \hat{\mu}_i)^2}{2\sigma^2}\Big) \\ \text{subject to} & |\hat{\mu}_i - \hat{\mu}_j| \leq K||X_i - X_j||_\infty, \qquad i, j = 1, 2, \ldots, n. \end{array}$$

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

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

We recognize this problem as the one described in Section 2. There is a clear 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.

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 (5), 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 $\hat{\mu}_1, \ldots, \hat{\mu}_n, \hat{\sigma}_1, \ldots, \hat{\sigma}_n$:

$$\begin{array}{lll} \text{maximize} & \displaystyle\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\hat{\sigma}_i} \exp\Big(-\frac{(Y_i - \hat{\mu}_i)^2}{2\hat{\sigma}_i^2}\Big) & \\ \text{subject to} & |\hat{\mu}_i - \hat{\mu}_j| \leq K_1||X_i - X_j||_\infty, & i, j = 1, 2, \ldots, n, \\ & |\hat{\sigma}_i - \hat{\sigma}_j| \leq K_2||X_i - X_j||_\infty, & i, j = 1, 2, \ldots, n. \end{array}$$

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

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

Note that here the objective function is not convex.

4.3 Bernoulli random variables

Suppose that we observe a sequence of binary random variables $Y_1, \ldots, Y_n$. Assume that $p(X_i) \equiv \Pr(Y_i = 1 \mid X_i)$ depends continuously on some observable variable X_i. In particular, the function $p : \mathcal{X} \mapsto [0,1]$ is Lipschitz continuous, with *unknown* constant K_0. Using the maximum likelihood approach (5), we may construct an estimated function $\hat{p}$ based on observations (X_1, Y_1),. . .,(X_n, Y_n), by solving the following optimization problem in the variables $\hat{p}_1, \ldots, \hat{p}_n$:

$$\begin{array}{ll} \text{maximize} & \displaystyle\prod_{i=1}^{n} \hat{p}_i^{Y_i}(1-\hat{p}_i)^{1-Y_i} \\ \text{subject to} & |\hat{p}_i - \hat{p}_j| \leq K||X_i - X_j||_\infty, \qquad i,j = 1,2,\ldots,n. \end{array}$$

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

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{i=1}^{n} Y_i \log(\hat{p}_i) + \sum_{i=1}^{n}(1-Y_i)\log(1-\hat{p}_i) \\ \text{subject to} & |\hat{p}_i - \hat{p}_j| \leq K||X_i - X_j||_\infty, \qquad i,j = 1,2,\ldots,n. \end{array}$$

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

4.4 Exponentially distributed random variables

Suppose that we observe a sequence of random values $Y_1, \ldots, Y_n$. We assume that Y_i is 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 (5), we may construct an estimated function $\hat{\lambda}$ based on observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, by solving the following optimization problem in the variables $\hat{\lambda}_1, \ldots, \hat{\lambda}_n$:

$$\begin{array}{ll} \text{maximize} & \displaystyle\prod_{i=1}^{n} \hat{\lambda}_i \exp(-\hat{\lambda}_i Y_i) \\ \text{subject to} & |\hat{\lambda}_i - \hat{\lambda}_j| \leq K||X_i - X_j||_\infty, \qquad i,j = 1,2,\ldots,n. \end{array}$$

Again by taking the logarithm, this is equivalent to

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{i=1}^{n} \log \hat{\lambda}_i - \sum_{i=1}^{n} \hat{\lambda}_i Y_i \\ \text{subject to} & |\hat{\lambda}_i - \hat{\lambda}_j| \leq K||X_i - X_j||_\infty, \qquad i,j = 1,2,\ldots,n. \end{array}$$

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

5 Simulation results

In this section, we provide some simulation results involving the Regression Algorithm from Section 2. We also compare its performance with kernel regression, on the same samples of artificially generated data.

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

$$Y = \sin X + \psi,$$

where $0 \leq X \leq 2\pi$ and the noise term ψ is normally distributed as $N(0, \sigma^2)$. We divide the interval $[0, 2\pi]$ into $n-1$ equal intervals and let $X_i = 2\pi(i-1)/(n-1)$, $i = 1, \ldots, n$, be the endpoints of the latter intervals. We generate n independent noise terms $\psi_1, \psi_2, \ldots, \psi_n$, with normal $N(0, \sigma^2)$ distribution and let $Y_i = \sin X_i + \psi_i$. We run Step 1 of the Regression Algorithm based on the pairs (X_i, Y_i), $i = 1, 2, \ldots, n$, and obtain the estimates $\hat{f}_1, \ldots, \hat{f}_n$. We also compute kernel regression estimates of the function $\sin x, x \in [0, 2\pi]$ using the same samples (X_i, Y_i). For the estimated functions $\hat{f}$ obtained by either the Regression Algorithm or kernel regression, we consider the performance measures

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

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 $\max_{0 \leq x \leq 2\pi} |\hat{f}(x) - \sin x|$ 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 $\max_{0 \leq x \leq 2\pi} |\hat{f}(x) - f(x)|$ for any Lipschitz continuous function $f(x)$. The second performance measure approximates the distance between $\sin x$ and $\hat{f}(x)$ with respect to the L_2 norm.

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 $\sigma = .5$. A Lipschitz constant $K = 2$ is used for this experiment. The piecewise linear curve around the curve $\sin(x)$ is the resulting estimated function $\hat{f}$. The points indicated by stars are the actual observations (X_i, Y_i), $i = 1, 2, \ldots, 100$. We see that the algorithm is successful in obtaining a fairly accurate approximation of the function $\sin x$.

In Tables 1 and 2, we summarize the results of several experiments, for the performance measures d_∞ and d_2, respectively. In all cases, the sample size is $n = 100$. 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, for which $|\sin(x) - \sin(y)| \leq K|x - y|$ 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.

$n = 100$	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 1. Experimental results with respect to the performance measure d_∞

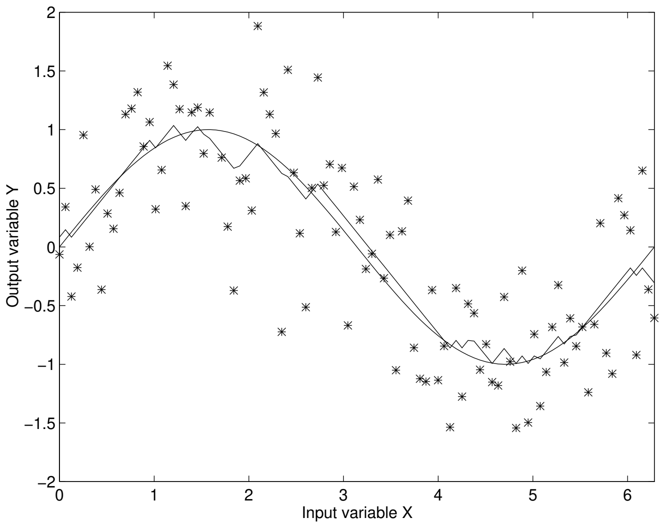

Figure 1: Experimental results with the Regression Algorithm, with $n = 100$, $\sigma = 0.5$ and $K = 1$

$n = 100$	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 2. Experimental results with respect to the performance measure d_2

The metric d_∞ is a more conservative measure of accuracy than the metric d_2. Therefore, it is not surprising that the approximation errors in Table 2 are larger. Examining the performance of the Regression Algorithm for the choices $K = 1$ and $K = 2$, we see that it is not particularly sensitive to the choice of K. The values obtained with $K = 1$ and $K = 2$ are quite close to each other. The dependence of the error on K is further demonstrated in Figure 2. We have computed the errors for samples of size $n = 50$ and constant K ranging from 0 to 10. Note that the optimal value is somewhat smaller than the correct one $K_0 = 1$, suggesting that it pays to somewhat underestimate the constant for the benefit of fewer degrees of freedom. Note that for large K we essentially overfit the data. Also the case $K = 0$ corresponds simply to a sample average.

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 ($\sigma = 0.5$ for $\delta = 0.3$; $\sigma = 0.1$ or $\sigma = 0.05$ 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.

Finally, we have investigated the dependence of the error d_2 on the sample size n. The results are reported in Figure 3. For every n between 10 and 100, we repeat the experiment 40 times, with $\sigma = .5$ We take the average squared error d_2^2 over these 40 experiments, and plot its negative logarithm. We also show the graphs of $\log(n)$ and $\log(n^{2/3})$ (shifted vertically, so that initial points coincide).

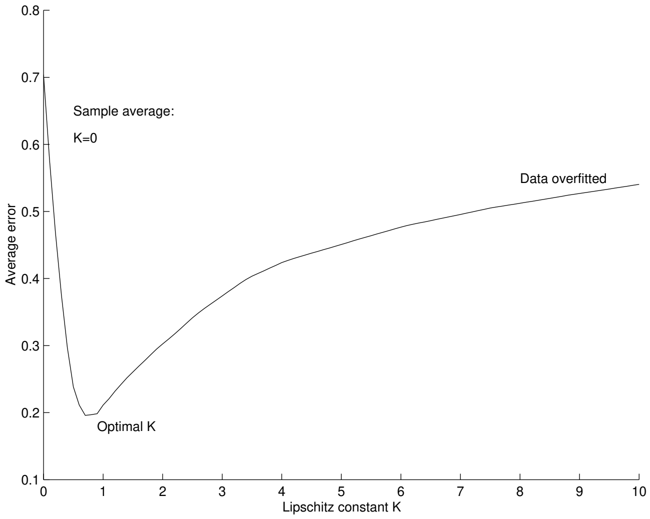

Figure 2: Approximation error as a function of K, with respect to the metric d_2.

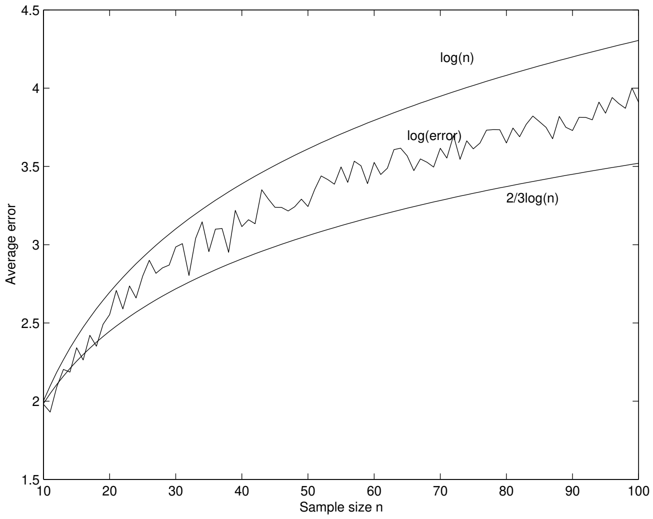

Figure 3: Plot of the negative logarithm of the squared error d_2^2 as a function of the sample size.

6 Convergence to the true regression function; consistency result.

In this section, we discuss the consistency of our convex optimization regression algorithm. Roughly speaking, we show that for the nonlinear regression model $Y = f(X) + \psi$ of 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 larger than the true constant K_0. Note that the boundedness assumption does not allow for, say, Gaussian noise and does not cover problems such as the one considered in Example 4.1. For any continuous scalar function g defined on the unit cube $[0,1]^d$, let the norm $||g||_\infty$ be defined as $\max_{x\in[0,1]^d} |g(x)|$.

Theorem 2 *Consider bounded random variables X, Y, with ranges in $[0,1]^d$ and $[0,1]$, respectively. Let $F(x,y)$ denote their joint probability distribution function. 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 density function $\phi(x)$, satisfying $\phi(x) \geq \beta > 0$ for all $x \in [0,1]^d$ and some $\beta > 0$.*

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 estimated function computed by the Regression Algorithm of Section 2. If $K \geq K_0$, 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 positive constants $\gamma_1(\epsilon)$ and $\gamma_2(\epsilon)$ such that*

$$\Pr\{||\hat{f}^n - f||_\infty > \epsilon\} \leq \gamma_1(\epsilon) e^{-\gamma_2(\epsilon)n}, \qquad \forall\ n. \tag{6}$$

Remarks:

1. Part 2 of the theorem implies that $\Pr\{||\hat{f}^n - f||_\infty > \epsilon\}$ can be made smaller than any given $\delta > 0$, by choosing n large enough. Explicit estimates for $\gamma_1(\epsilon)$ and $\gamma_2(\epsilon)$ are readily obtained from the various inequalities established in the course of the proof. These estimates are too conservative to be practically useful. The estimates also indicate that the number of required samples increases exponentially with the dimension d, but this is unavoidable, even in the absence of noise.

2. Theorem 2 can be easily extended to the case where the range of the input variable X is some rectangle $\prod_{i=1}^d [a_1^l, a_2^l]$, and the range of the output variable Y is some interval $[b_1, b_2]$. The extension is obtained by rescaling the input and output variables.

Proof: Let $\Im$ be the set of all Lipschitz continuous functions $\hat{f} : [0,1]^d \mapsto [0,1]$ with constant K. Introduce the risk function

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

defined on $[0,1]^{d+1} \times \Im$. The estimate $\hat{f}^n$ obtained from steps 1 and 2 of the Regression Algorithm is a solution to the problem

$$\text{minimize} \quad \frac{1}{n}\sum_{i=1}^n Q(X_i, Y_i, \hat{f}) \qquad \text{over } \hat{f} \in \Im. \tag{7}$$

In particular

$$\sum_{i=1}^n Q(X_i, Y_i, \hat{f}^n) \leq \sum_{i=1}^n Q(X_i, Y_i, f). \tag{8}$$

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

$$\text{minimize} \int Q(x,y,\hat{f})\, dF(x,y) \qquad \text{over } \hat{f} \in \Im,$$

because for any fixed $x \in [0,1]^d$, the minimum of $E[(Y - \hat{f}(x))^2 | X = x]$ is achieved by $\hat{f}(x) = E[Y \mid X = x] = f(x)$.

Our proof of Theorem 2 is built on the concept of *VC entropy* [1]. For any given set of pairs

$$(x_1, y_1), \ldots, (x_n, y_n) \in [0,1]^{d+1}$$

consider the set of vectors in $\Re^n$

$$\{(Q(x_1, y_1, \hat{f}), \ldots, Q(x_n, y_n, \hat{f})) : \ \hat{f} \in \Im\} \tag{9}$$

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 $q_1, q_2, \ldots, q_k \in \Re^n$, such that for any vector q in the set (9), $||q - q_j||_\infty < \epsilon$ for some $j = 1, 2, \ldots, k$. The following definition of VC entropy was used by Haussler in [6].

Definition 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 was proven by Lee, Bartlett and Williamson [8]. It improves on earlier results by Pollard (Theorem 24, page 25, [5]) and Haussler [6].

Proposition 2 *For every $\alpha > 0$, we have*

$$\Pr\left\{ \sup_{\hat{f} \in \Im} \Big| (1-\alpha)\Big(\int Q(x,y,\hat{f}) dF(x,y) - \int Q(x,y,f) dF(x,y) \Big) - \Big(\frac{1}{n}\sum_{i=1}^n Q(X_i, Y_i, \hat{f}) - \frac{1}{n}\sum_{i=1}^n Q(X_i, Y_i, f) \Big) \Big| > \alpha^2 \right\}$$
$$\leq 6 H^{\Im}\Big(\frac{\alpha^2}{256}, n\Big) e^{\frac{-3\alpha^3 n}{5248}}.$$

Remark: This bound is readily obtained from Theorem 3 of [8], by setting $\nu = \nu_c = \alpha/2$.

The key to our analysis is to show that for the 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, and admits a bound that does not depend on the sample size n.

Proposition 3 *For any $\epsilon > 0$ and any sequence $(x_1, y_1), \ldots, (x_n, y_n)$ in $[0,1]^{d+1}$, there holds*

$$N(\epsilon, \Im, (x_1, y_1), \ldots, (x_n, y_n)) \le \Big(\frac{4}{\epsilon} + 1\Big) 2^{\frac{(2K)^d}{\epsilon^d}}.$$

In particular,

$$H^{\Im}(\epsilon, n) \le \Big(\frac{4}{\epsilon} + 1\Big) 2^{\frac{(2K)^d}{\epsilon^d}}$$

and

$$\log H^{\Im}(\epsilon, n) = O\Big(\Big(\frac{K}{\epsilon}\Big)^d\Big).$$

and the bound on the VC entropy does not depend on n.

This result is based on a theorem by Kolmogorov and Tihomirov [7] on the VC entropy (ϵ-covering number) of the space of Lipschitz continuous functions. We provide a statement of this theorem and a proof of Proposition 3 in the Appendix.

For any function $g \in \Im$, its L_2-norm $||g||_2$ is defined by

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

In the following proposition, we obtain a bound on the tail probability of the difference $||\hat{f}^n - f||_2$.

Proposition 4 *There holds*

$$\Pr\Big\{||\hat{f}^n - f||_2 > \epsilon\Big\} \le 6H^{\Im}\Big(\frac{\epsilon^2}{2^{10}}, n\Big) e^{\frac{-3\epsilon^3 n}{41 \times 2^{10}}}. \tag{10}$$

for all $\epsilon < 1$.

Proof: See the Appendix.

Combining Propositions 3 and 4, we immediately obtain the following result.

Proposition 5 *There holds*

$$\Pr\Big\{||\hat{f}^n - f||_2 > \epsilon\Big\} \le \Big(\frac{2^{12}}{\epsilon^2} + 1\Big) 2^{\frac{2^{11d} K^d}{\epsilon^{2d}}} e^{\frac{-3\epsilon^3 n}{41 \times 2^{10}}}. \tag{11}$$

for all $\epsilon < 1$.

Our next step is to show that $||\hat{f}^n - f||_\infty \to 0$ almost surely. The following lemma establishes that convergence in $||\cdot||_2$ norm implies the convergence in $||\cdot||_\infty$ for the class $\Im$ of Lipschitz continuous functions with constant K. This will allow us to prove a result similar to (11) but with $||\hat{f}^n - f||_2$ replaced by $||\hat{f}^n - f||_\infty$.

Lemma 3 *Consider a Lipschitz continuous function $g : [0,1]^d \to \Re$ with Lipschitz constant K. Suppose that for some $\epsilon > 0$ there holds $||g||_\infty \ge \epsilon$. Then,*

$$||g||_2 \ge \frac{\epsilon^{\frac{d}{2}+1} \beta^{\frac{1}{2}}}{2^{\frac{d}{2}+1} K^{\frac{d}{2}}}.$$

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 $a \in [0,1]^d$, we have $|g(a)| \geq \epsilon$. Set $\delta = \epsilon/(2K)$. We have

$$||g||_2^2 \geq \int_{x:||x-a||_\infty \leq \delta} g^2(x)dF(x)$$

For any x such that $||x-a||_\infty \leq \delta$ we have $|g(x) - g(a)| \leq K\delta$. It follows that $|g(x)| \geq \epsilon - K\delta = \epsilon/2$, whenever $||x-a||_\infty \leq \delta$. In the integral above, we are only integrating over elements of the unit cube that satisfy $\|x-a\|_\infty \leq \delta$. In the worst case, where a is a corner point, we are integrating over a set of volume δ^d. Furthermore, the density is at least β. Therefore,

$$||g||_2^2 \geq \frac{\epsilon^2}{4}\Pr\Big\{||X-a||_\infty \leq \frac{\epsilon}{2K}\Big\} \geq \frac{\epsilon^2}{4}\beta\frac{\epsilon^d}{(2K)^d} > 0,$$

and the result follows by taking square roots. □

Lemma 3 implies that

$$\Pr\Big\{||\hat{f}^n - f||_\infty > \epsilon\Big\} \leq \Pr\Big\{||\hat{f}^n - f||_2 > \frac{\epsilon^{\frac{d}{2}+1}\beta^{\frac{1}{2}}}{2^{\frac{d}{2}+1}K^{\frac{d}{2}}}\Big\}.$$

A bound for the right-hand side is provided by Proposition 5 and part 2 of the theorem follows immediately.

We have so far established the convergence of $||\hat{f}^n - f||_\infty \to 0$ in probability. 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 part 2 and the Borel-Cantelli lemma. □

The bounds established in the course of the proof provide 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 $\hat{f}^n = \hat{f}^n(\cdot;X_1,Y_1,\ldots,X_n,Y_n)$. Then given an arbitrary input observation $X \in [0,1]^d$ the probability that the deviation of the estimated output $\hat{f}^n(X)$ from the true output $f(X)$ is more than ϵ, is readily bounded above. Note, that the bound depends only on the distribution of X (through β) and not on the conditional distribution of Y given X. Unfortunately, the constants $\gamma_1(\epsilon)$ and $\gamma_2(\epsilon)$ are too large for practical purposes, even for dimension $d=1$. Our simulation results from Section 5 suggest that the rate of convergence of $\hat{f}^n$ to f is much better than predicted by our pessimistic bounds. It would be interesting to investigate whether better rates and more useful upper bounds can be established.

7 Extensions

As suggested by Theorem 2, the number of samples required to learn a Lipschitz continuous function can be huge when the input variable is multidimensional. This problem can be potentially overcome by making additional structural assumptions. For instance, assume that the input variable x can be represented as a pair of variables $(t,z), t \in \Re, z \in \Re^d$ (t could be time, for example). Assume that the regression function has the form $f(t,z) = b(t)'z$, where $b(t)$ is a Lipschitz continuous vector-valued function of a single variable, with Lipschitz constant K. Given such a model

$$Y = f(t,Z) + \psi$$

and a sequence of observations $(t_1,Z_1,Y_1),\ldots,(t_n,Z_n,Y_n)$, the following version of the regression algorithm provides an estimate of the underlying regression function f.

Step 1. Choose a constant K and solve the following constrained optimization problem in n (d-dimensional) variables $\hat{b}_1, \ldots, \hat{b}_n$:

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{n}(Y_i - \hat{b}_i' Z_i)^2 && (12)\\ &\text{subject to} && ||\hat{b}_i - \hat{b}_j||_\infty \leq K|t_i - t_j|, && i,j = 1,2,\ldots,n. \end{aligned}$$

Step 2. Let

$$\hat{f}(t,Z) = \Big(\frac{t - t_i}{t_{i+1} - t_i}\hat{b}_i' + \frac{t_{i+1} - t}{t_{i+1} - t_i}\hat{b}_{i+1}'\Big)Z.$$

for all $t \in [t_i, t_{i+1}]$.

Thus, in the first step the function $b(t)$ is estimated at the observed values (times) t_i, and then the values are extrapolated linearly for any values of t and X. When the value of the parameter K is small, we can think of this model as a linear regression model in which the regression vector is slowly varying with time. Note, that from Proposition 3, each of the coordinate functions $b_i(t), i = 1,2,\ldots,d$, has bounded VC entropy, which is independent from the dimension d and the sample size n. Therefore, the logarithm of the VC entropy of the space of Lipschitz continuous functions $b(t)$ is $O(d)$. This is better than $\Omega(2^d)$, which is known to be a lower bound for the case of general Lipschitz continuous functions of d variables.

A good approximation of the Lipschitz constant K is very important for the Regression Algorithm to be successful. If the input space $\mathcal{X}$ can be partitioned into, say, two parts $\mathcal{X}_1$,$\mathcal{X}_2$, with $\mathcal{X}_1 \cup \mathcal{X}_2 = \mathcal{X}$, such that within each part $\mathcal{X}_r$ a better estimate K_r of the constant K is available, such knowledge can be incorporated into the model (1) by using a tighter constraint $|\hat{f}_i - \hat{f}_j| \leq K_r||X_i - X_j||_\infty$, whenever $X_i, X_j \in \mathcal{X}_r, r = 1,2$.

We have observed in our simulation studies that the performance of the Regression Algorithm is comparable to kernel regression, for moderate magnitudes of noise variance. However, the kernel regression method has the advantage that the convergence rate, when $d = 1$, of the expected error $E[(\hat{f}^n - f)^2] = ||\hat{f}^n - f||_2^2$ is $O(n^{-2/3})$, which is the best possible. We have not proven a similar convergence rate for our Regression Algorithm. The best bound that can be obtained using the bounds on the tail probability $\Pr\{||\hat{f}^n - f||_2^2 > \epsilon\}$, is $O(n^{-2/5})$, and does not match the optimum. However, our simulations suggest that $O(n^{-2/3})$ is the right rate for our method as well. It is possible that a mixture of the two approaches produces a more desirable procedure. Such a mixture can be constructed as follows. Let $\phi(x_1,x_2), x_1,x_2 \in \mathcal{X}$ be the weight function used for kernel regression. Thus, given a sequence of observations $(X_i,Y_i), i = 1,2,\ldots,n$, kernel regression produces the estimates

$$\hat{f}(x) = \frac{\sum_{i=1}^{n}\phi(x,X_i)Y_i}{\sum_{i=1}^{n}\phi(x,X_i)}.$$

Note, that the resulting values $\hat{f}(X_i)$ can be viewed as solutions to the problem of minimizing

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\phi(X_i,X_j)(Y_i - \hat{f}_j)^2$$

with respect to the variables $\hat{f}_i$. If the underlying function is known to be Lipschitz continuous, this knowledge can be incorporated in additional constraints of the form

$$|\hat{f}_i - \hat{f}_j| \leq K||X_i - X_j||_\infty.$$

To what extent this mixed estimation procedure is advantageous over pure kernel regression or pure quadratic optimization is a subject for future research.

8 Conclusions

We have proposed a convex optimization approach to the nonparametric 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 the consistency properties of this estimation procedure. Other questions for further investigation relate to the bounds on the expected error $(E[||\hat{f}^n - f||_2^2])^{\frac{1}{2}}$ and to methods for setting a value for the constant K. A good choice of K is crucial for the approximation to be practically successful.

Acknowledgments. We would like to thank the anonymous referees for many constructive suggestions on how to improve both the substance and the presentation in this paper.

9 Appendix

We provide in this appendix the proofs of Propositions 3 and 4.

Proof of Proposition 3:

Kolmogorov and Tihomirov ([7], p. 356) proved the following theorem.

Theorem 4 *Let $\Im_1$ be the space of Lipschitz continuous functions defined on the unit cube $[0,1]^d$, bounded by some constant B, and having Lipschitz constant $K = 1$. Then the size $N(\epsilon, \Im_1)$ of the minimal ϵ net of $\Im_1$ satisfies*

$$N(\epsilon, \Im_0) \le \left(2\left\lfloor\frac{2B}{\epsilon}\right\rfloor + 1\right) 2^{\frac{1}{\epsilon^d}}.$$

Consider our set $\Im$ of Lipschitz continuous functions with range $[0,1]$. By dividing all of these functions by K and subtracting $1/(2K)$, we obtain the set of Lipschitz continuous functions with range $[-1/(2K), 1/(2K)]$. Applying Theorem 4, the minimal size of an (ϵ/K)-net in this set is no larger than

$$\left(2\left\lfloor\frac{\frac{2}{(2K)}}{\frac{\epsilon}{K}}\right\rfloor + 1\right) 2^{\frac{1}{(\frac{\epsilon}{K})^d}} = \left(2\left\lfloor\frac{1}{\epsilon}\right\rfloor + 1\right) 2^{\frac{K^d}{\epsilon^d}}.$$

It follows that the minimal size $N(\epsilon, \Im)$ of the ϵ-net of the set $\Im$ satisfies

$$N(\epsilon, \Im) \le \left(\frac{2}{\epsilon} + 1\right) 2^{\frac{K^d}{\epsilon^d}}.$$

To complete the proof of Proposition 3, we relate the minimal ϵ-net size of $\Im$ to the minimal ϵ-net size $N(\epsilon, \Im, (x_1, y_1), \ldots, (x_n, y_n))$, of the set

$$\{(Q(x_1, y_1, f), \ldots, Q(x_n, y_n, f),\ f \in \Im\} = \{(y_1 - f(x_1))^2, \ldots, (y_n - f(x_n))^2,\ f \in \Im\}. \tag{13}$$

For any two functions $f, g \in \Im$ and any $i = 1, 2, \ldots, n$, we have

$$|(y_i - f(x_i))^2 - (y_i - g(x_i))^2| = |f(x_i) - g(x_i)| \cdot |2y_i - f(x_i) - g(x_i)| \le 2|f(x_i) - g(x_i)|.$$

It follows, that for any ϵ the minimal size $N(\epsilon, \Im, (x_1, y_1), \ldots, (x_n, y_n))$ of an ϵ-net of the set (13) is at most

$$\Big(\frac{4}{\epsilon}+1\Big)2^{\frac{(2K)^d}{\epsilon^d}}.$$

This completes the proof.□

Proof of Proposition 4: The identity

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

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

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

and

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

where the last equality is a consequence of $E[Y \mid X] = f(X)$. Then,

$$\begin{aligned} & \Pr\Big\{\sup_{\hat{f}\in\Im}\Big|(1-\alpha)\Big(\int Q(x,y,\hat{f})dF(x,y) - \int Q(x,y,f)dF(x,y)\Big) \\ & \qquad -\Big(\frac{1}{n}\sum_{i=1}^{n} Q(X_i,Y_i,\hat{f}) - \frac{1}{n}\sum_{i=1}^{n} Q(X_i,Y_i,f)\Big)\Big| > \alpha^2\Big\} \\ = \ & \Pr\Big\{\sup_{\hat{f}\in\Im}\Big|(1-\alpha)||\hat{f}-f||_2^2 - \Big(\frac{1}{n}\sum_{i=1}^{n} Q(X_i,Y_i,\hat{f}) - \frac{1}{n}\sum_{i=1}^{n} Q(X_i,Y_i,f)\Big)\Big| > \alpha^2\Big\} \\ \geq \ & \Pr\Big\{(1-\alpha)||\hat{f}^n-f||_2^2 - \Big(\frac{1}{n}\sum_{i=1}^{n} Q(X_i,Y_i,\hat{f}^n) - \frac{1}{n}\sum_{i=1}^{n} Q(X_i,Y_i,f)\Big) > \alpha^2\Big\} \\ \geq \ & \Pr\Big\{(1-\alpha)||\hat{f}^n-f||_2^2 > \alpha^2\Big\}, \end{aligned}$$

where the last inequality follows from (8). For all $\epsilon < 1$, we have $\epsilon^2 > \epsilon^2/(4(1-\epsilon/2))$. Therefore

$$\Pr\Big\{(1-\frac{\epsilon}{2})||\hat{f}^n-f||_2^2 > \frac{\epsilon^2}{4}\Big\} \geq \Pr\Big\{||\hat{f}^n-f||_2 > \epsilon\Big\}. \tag{15}$$

By setting $\alpha = \epsilon/2$, using (15) and (15), and applying Proposition 2, we obtain

$$\Pr\Big\{||\hat{f}^n-f||_2 > \epsilon\Big\} \leq 6H^{\Im}\Big(\frac{\epsilon^2}{2^{10}}, n\Big)e^{\frac{-3\epsilon^3 n}{41\times 2^{10}}}.$$

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 Transactions on Information Theory,* 24, 142-151, 1978.

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

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

[6] D.Haussler. Decision Theoretic Generalization of the PAC model for Neural Net and Other Learning Applications. *Information and Computation*, 100, 78-150, 1992.

[7] A.N. Kolmogorov and V.M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *American Mathematical Translations*, 17, 277-364, 1961.

[8] W. Lee, P. Bartlett and R. Williamson. Efficient Agnostic Learning of Neural Networks with Bounded Fan-in. *IEEE Transactions on Information Theory*, 42:6, 2118-2132, 1996.