Efficient Learning of Monotone Concepts via Quadratic Optimization

David Gamarnik *

October 14, 1999

Abstract

We consider a non-parametric regression model, in which the regression function to be estimated is coordinate-wise monotone. Such functions can model, for example, the dependence of an option price on a strike price, duration and a price of an underlying asset. We propose a simple algorithm based on quadratic optimization, which estimates the regression function in polynomial time. Our main theoretical result shows that the expected VC-entropy of the space of coordinate-wise monotone functions grows superpolynomially but subexponentially. As a result, the estimating function, constructed by our algorithm, converges to the true regression function, as the number of observations goes to infinity, and the estimating procedure is consistent. Numerical results are provided, which show that the algorithm is quite robust to possible discontinuities of the regression function.

1 INTRODUCTION

We consider a regression model of the form

$$Y = f(X) + \psi,$$

where X, Y is an observable pair of input-output random variables and ψ is a zero mean unobserved random noise (in particular $E[Y|X] = f(X)$). The regression function $f : \mathcal{D} \to \Re$ is defined on some bounded domain $\mathcal{D} \subset \Re^d$ and is coordinate-wise monotone. Namely, for each $i = 1, 2, \ldots, d$ f is a non-decreasing or non-increasing function of the i-th coordinate, when all the other coordinates are fixed. If the regression function f is differentiable, this simply means all partial derivatives preserve the signs throughout the domain. However, we do allow in our model for f to be discontinuous. The main problem is to estimate a function f based on a sequence of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$.

*Mathematical Sciences Department IBM, T.J.Watson Research Center P.O.Box 218, Yorktown Heights, NY 10598 gamarnik@watson.ibm.com

Monotone regression functions can model a variety of industrial, financial or biological processes. For example, the number of customers in a bank or number of calls in progress in a telephone network at time t both can be viewed as non-decreasing one dimensional function $f(t)$ of time t, within certain time intervals of the day (say between 7am and 12 pm). By observing a number of customers or calls at certain times $t_1 < t_2 < \cdots < t_n$ within the interval, we obtain noisy observations $Y_1, Y_2, \ldots, Y_n$ of the regression function f. A multidimensional example of a monotone regression model is option pricing in financial markets. The price of an American call option is an increasing function of the duration and the price of the underlying asset, and a decreasing function of the strike price. Under certain conditions on dynamics of securities (stocks and bonds) the price of the corresponding option can be calculated in closed form using the well-known Black-Scholes formula. However in general the calculation of the price is problematic. An alternative approach was considered in by Hutchinson, Lo and Poggio in [2], where statistical regression was used to learn the price from market observations. Another example of a multidimensional monotone regression model is, for instance, a price of a house as an increasing function of the size and decreasing function of the distance from downtown.

The literature on regression models with monotone properties is not comprehensive. Kearns and Schapire considered in [3] a one dimensional monotone model, in which the output variables take values in $\{0, 1\}$. The regression function f was called monotone probabilistic concept. The statistical regression used in [2] were based on Radial Basis Functions, Multilayer Perceptron and Projection Pursuit, and did not explicitely capitalize on the monotonicity property of the option price function.

In this paper we show how quadratic optimization technique can be used for constructing an estimate function $\hat{f}$ of the monotone regression function f, based on a sequence of observations of input-output variables $(X_1, Y_1), \ldots, (X_n, Y_n)$. The quadratic optimization technique as a method of learning regression functions was first proposed in [1] for learning Lipschitz continuous concepts. This paper can be viewed as an extension of these results to monotone concepts.

In section 2 we propose an polynomial time algorithm for estimating function f. The algorithm is based on a certain quadratic optimization problem with linear constraints. The objective function is a sum of squared errors between the estimated and observed values. A set of linear constraints specifies the monotonicity property of f. A certain estimating function $\hat{f}^n = \hat{f}^n(X_1, Y_1, \ldots, X_n, Y_n)$ is constructed as a result of this procedure.

Our main theoretical results are summarized in Section 3. We consider a particular case of a two dimensional domain $[0, 1] \subset \Re^2$ and assume that the distribution of the input variables $X_1, X_2, \ldots, X_n$ has continuous positive density. A notion of the expected *VC-entropy* is used to study the asymptotic properties of our constructed function $\hat{f}^n$, when n diverges to infinity. We prove that the expected VC-entropy diverges to infinity at the rate $\Theta(e^{\theta(n^{\frac{1}{2}})})$. In particular, the growth rate is superpolynomial and subexponential. Then, using results of Vapnik and Chervonenkis, Pollard and Haussler, we prove that $\hat{f}^n$ converges to f in L_2 norm, as n diverges to infinity. As a result, our estimating procedure is consistent.

The algorithm is applied in Section 5 to estimating a time varying parameter of non-stationary Poisson process from an observed sequence of arrivals. The simulation results in Section 4 show that our algorithm is very robust to discontinuities of the regression function f. It is also robust to different levels on the noise. In particular, when the noise level is very small, our algorithm outperforms kernel regression method by orders of magnitude, and is comparable to or better than kernel regression method, when the noise level is medium.

2 A MONOTONE REGRESSION MODEL AND THE ALGORITHM

We consider a regression model

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

where X and Y are a pair of input-output random variables with ranges in $\Re^d$ and $\Re$ respectively, and $\Re^d$ is a usual d-dimensional Euclidean space. Noise ψ is assumed to be a zero-mean random variable.

The regression function $f : \mathcal{D} \to \Re$ is some unknown function, defined on a bounded domain $\mathcal{D} \subset \Re^d$. The function f is assumed to be coordinate-wise monotone in the following sense. There exists an a priori known d-dimensional vector $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_d)$, with coordinates equal to 1 or -1, with the following property. If $x = (x_1, \ldots, x_d), y = (y_1, \ldots, y_d)$ are two vectors from $\mathcal{D}$, satisfying $x_i \le y_i$ when $\sigma_i = 1$ and $x_i \ge y_i$, when $\sigma_i = -1$, then $f(x) \le f(y)$. In words, f is a non-decreasing function across the coordinates i, with $\sigma_i = 1$, and is a non-increasing function across the coordinates i, with $\sigma_i = -1$. If f has partial derivatives, then coordinate-wise monotonicity simply means partial derivatives preserve the signs on the domain $\mathcal{D}$. The model (1) will be referred to as *monotone regression model*. Note that by substituting all the variables x_i with $-x_i$, for all i with $\sigma_i = -1$ (and changing respectively the domain $\mathcal{D}$) we may assume without the loss of generality that all the components of σ are equal to one. Namely, f is a non-decreasing function across all the coordinates. Simply putting

Assumption. The function f is assumed to be coordinate-wise non-decreasing: for any $x, y \in \mathcal{D}$ $f(x) \le f(y)$ if $x_i \le y_i$ for all $i = 1, 2, \ldots, d$.

Note, in a one dimensional case, this simply means, f is a non-decreasing function. The condition $x_i \leq y_i, i = 1, 2, \ldots, d$ defines a partial order on the set $\Re^d$, which we denote by $\prec$. Given model (1) and a sequence of observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, our objective is to construct an estimate $\hat{f} = \hat{f}^n(X_1, Y_1, \ldots, X_n, Y_n)$ of the function f, such that $\hat{f}^n$ is also defined on the domain $\mathcal{D}$, and is also coordinate-wise non-decreasing. As a solution we propose the following two-step algorithm.

Monotone Regression Algorithm

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

$$\text{minimize} \sum_{i=1}^{n} (Y_i - \hat{f}_i)^2 \tag{2}$$

subject to

$$\hat{f}_i - \hat{f}_j \leq 0 \text{ for all } i, j \text{ satisfying } X_i \prec X_j. \tag{3}$$

This step gives predictions $\hat{f}_i \equiv \hat{f}(X_i)$ of the correct outputs $f(X_i)$ at the input points $X_1, X_2, \ldots, X_n$.

Step 2. In this step, we extrapolate the values $\hat{f}_1, \ldots, \hat{f}_n$ obtained in Step 1, to a coordinate-wise non-decreasing function $\hat{f} : \mathcal{D} \to \Re$ as follows: for any $x \in \mathcal{D}$, let

$$\hat{f}(x) = \max_{1 \leq i \leq n} \{\hat{f}_i : X_i \prec x\}.$$

Namely, for each point x we assign the smallest value, which is consistent with values of $\hat{f}$ at all the points X_i smaller than x. If none of the points $X_1, X_2, \ldots, X_n$ is smaller than x, then we either put $\hat{f}(x)$ equal to $-\infty$ or equal to any a priori known lower bound of $\inf_{x \in \mathcal{D}}\{f(x)\}$. Note, that this is the best we can do when none of the points X_i is smaller than x, since f is not assumed to be continuous and in principle can take any of the values in the interval $(-\infty, \min_{1 \leq i \leq n}\{\hat{f}_i : x \prec X_i\}]$.

In the following proposition we justify the Step 2 of the algorithm.

Proposition 1 *The function $\hat{f}$ defined above is a coordinate-wise non-decreasing function. It satisfies*

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

Proof: For each $k = 1, 2, \ldots, n$ we have $\hat{f}(X_k) = \max_{1 \leq i \leq n}\{\hat{f}_i : X_i \prec X_k\}$. From constraints (3) $\hat{f}_i \leq \hat{f}_k$ when $X_i \prec X_k$, therefore $\hat{f}(X_k) = \hat{f}_k$.

Let $x, y \in \mathcal{D}$ satisfy $x \prec y$. Then $X_i \prec x$ implies $X_i \prec y$ and

$$\hat{f}(x) = \max_{1 \leq i \leq n} \{\hat{f}_i : X_i \prec x\} \leq \max_{1 \leq i \leq n} \{\hat{f}_i : X_i \prec y\} = \hat{f}(y) \; \Box.$$

Note that in Step 2 of the algorithm we could take instead $\hat{f}(x) = \min_{1\le i\le n}\{\hat{f}_i : x \prec X_i\}$ or $\hat{f}(x) = (1/2)\min_{1\le i\le n}\{\hat{f}_i : x \prec X_i\} + (1/2)\max_{1\le i\le n}\{\hat{f}_i : X_i \prec x\}$.

Remarks:

1. As we see, the algorithm is designed to minimize the sum of empirical squared errors, subject to the monotonicity constraints. Notice, that when the dimension d is bigger than one, the number of the constraints can take any value between zero and $(1/2)n(n-1)$. We will be assuming that the input values are drawn i.i.d from a certain distribution. In that case, as we will see in Section 3, reasonably many points will be comparable, with a certain positive probability. When the dimension d is equal to one, the number of constraints is trivially reduced to $n-1$.

2. Optimization problem (2),(3) can be solved efficiently, since the objective function is quadratic (convex), all the constraints are linear and the total number of constraints is at most $(1/2)n(n-1)$ (see [7]).

3. Our algorithm is an extension of the Regression Algorithm proposed in [1] for learning Lipschitz continuous regression functions f. There, an a priori assumption of Lipschitz continuity was explicitely put as a constraint in the optimization problem. Here we put monotonicity as an explicit constraint in the optimization problem.

4. We emphasize the fact that f is not assumed to be a continuous function, and the algorithm does not assume the knowledge of the points of discontinuities. As our numerical results suggest, the algorithm is very robust to discontinuities. Also note, that if the noise terms $\psi_i, i = 1, 2, \ldots, n$ are all equal to 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 other non-parametric regression methods like kernel regression or smoothing splines, where due to the selected positive bandwidth (kernel regression) or regularization parameter (smoothing splines), the estimating function is not equal to the true function, even if the noise is zero.

3 EXPECTED VC-ENTROPY AND CONVERGENCE RESULTS

In this section we investigate the consistency properties of our monotone regression algorithm - does an estimating function $\hat{f}^n$ converge to the true regression function f, when the number of observations n diverges to infinity? We establish certain consistency results, by incorporating a concept of *expected VC entropy*, introduced by Vapnik and Chervonenkis (see [6]).

Let $\Im$ be some family of functions $\hat{f} : \mathcal{D} \to [b_1, b_2] \subset \Re$ defined on a bounded domain $\mathcal{D} \subset \Re^n$. Let $Q : \mathcal{D} \times [b_1, b_2] \times \Im \to \Re_+$ be a non-negative cost function defined as follows $Q(x, y, f) = (y - \hat{f}(x))^2$, for $x \in \mathcal{D}, y \in [b_1, b_2], \hat{f} \in \Im$. Note, that if $\Im$ is the space of coordinate-wise monotone functions, then the Monotone Regression Algorithm terminates in a solution $\hat{f}$ to the following optimization problem

$$\text{Minimize}_{\hat{f} \in \Im} \sum_{i=1}^{n} Q(X_i, Y_i, \hat{f}). \tag{4}$$

For any fixed sequence $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{D} \times [b_1, b_2]$ consider the set of vectors

$$\{(Q(x_1, y_1, \hat{f}), \ldots, Q(x_n, y_n, \hat{f})),\ \hat{f} \in \Im\} \tag{5}$$

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 (5), $||q - q_j||_\infty < \epsilon$ for some $j = 1, 2, \ldots, k$, where $||\cdot||_\infty$ is the maximum norm in $\Re^n$. Suppose $H(x, y)$is a certain probability distribution function $H(x, y)$ on the space $\mathcal{D} \times [b_1, b_2]$, and $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are chosen independently and randomly according to $H(x, y)$. The following definition of (expected) VC entropy was used by Haussler in [8].

Definition 1 *For any $\epsilon > 0$, the VC entropy of $\Im$ for samples of size n is*

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

where the expectation is with respect to the distribution $H(x, y)$.

The (expected) VC entropy $N^{\Im}(\epsilon, n)$ measures, in a certain sense, the number of degrees of freedom of a concept space $\Im$, on an input vector $X_1, X_2, \ldots, X_n$. Vapnik and Chervonenkis have proven, that the least squares estimates lead to a consistent estimation, when $N^{\Im}(\epsilon, n)$ diverges to infinity at most subexponentially, namely

$$\frac{\log(N^{\Im}(\epsilon, n))}{n} \to 0, \tag{6}$$

as $n \to \infty$. The following theorem is proven in [1]. It is obtained as a direct corollary of the bound (6) and its refinements obtained subsequently by Pollard [5], Haussler [8] and [4]. Let $H(x)$ denote the marginal distribution of the input variable X, and let $||\cdot||_2$ denote the L_2 norm with respect to measure $H(x)$. Namely,

$$||g||_2^2 = \int_{\mathcal{D}} g^2(x) dH(x) = E[g^2(X)].$$

Proposition 2 *For every* $\epsilon < 1$

$$\Pr\Big\{||\hat{f}^n - f||_2 > \epsilon\Big\} \leq 6H^{\Im}\Big(\frac{\epsilon^2}{2^{10}}, n\Big) e^{\frac{-3\epsilon^3 n}{41\times 2^{10}}}. \tag{7}$$

Proof: See the proof of Proposition 4 in [1]

Thus, in order to prove the consistency of our algorithm, we need to show that the expected VC-entropy corresponding to a family of coordinate-wise monotone functions diverges at most subexponentially as a function of n, that is (6) holds. We will prove this when $\mathcal{D}$ is either a one-dimensional interval or a two-dimensional square, and when the marginal distribution $H(x)$ of the input variable X has a continuous and positive density on $\mathcal{D}$.

In order to estimate the expected VC entropy of the family of coordinate-wise monotone functions, we relate it to a certain combinatorial problem, introduced as follows

Labeling number. Given a sequence of points $x_1, \ldots, x_n$ from a bounded region $\mathcal{D} \subset \Re^n$ and a positive integer m, let $L(m, x_1, \ldots, x_n)$ denote the number of functions $\phi : \{x_1, \ldots, x_n\} \to \{1, 2, \ldots, m\}$ satisfying $\phi(X_i) \leq \phi(X_j)$ whenever $x_i \prec x_j$, $i, j = 1, 2, \ldots, n$. Namely, $L(m, x_1, \ldots, x_n)$ is the number of ways we can label the points $x_1, \ldots, x_n$ consistently with the partial order $\prec$. $L(m, x_1, \ldots, x_n)$ is defined to be the **labeling number** on the sequence. $x_1, x_2, \ldots, x_n$. Let also $L(m, n, H)$ denote the expected value of $L(m, X_1, \ldots, X_n)$, when $X_1, \ldots, X_n$ are drawn independently according to some distribution $H = H(x)$ on the domain $\mathcal{D}$.

Proposition 3 *Let* $\Im$ *denote the family of coordinate-wise monotone functions* $\hat{f} : \mathcal{D} \to [b_1, b_2]$. *Given two sequences* $x_i \in \mathcal{D}, y_i \in [b_1, b_2]$ $i = 1, 2, \ldots, n$ *and* $\epsilon > 0$, *the following bound holds*

$$N(\epsilon, \Im, (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \leq L(2\lceil \frac{b_2 - b_1}{\epsilon} \rceil, x_1, x_2, \ldots, x_n).$$

If, in addition $y_i \in \{b_1, b_2\}$, $i = 1, 2, \ldots, n$, *then*

$$N(\epsilon, \Im, (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \geq L(\lfloor \frac{b_2 - b_1}{\sqrt{2\epsilon}} \rfloor, x_1, x_2, \ldots, x_n).$$

Proof: See the Appendix.

We see that the labeling number an estimate on the VC entropy of the class of coordinate-wise monotone functions. We now state and prove the convergence result when $\mathcal{D}$ is a one dimensional interval $[a_1, a_2]$.

Theorem 1 *Consider bounded random variables $X, Y \in \Re$, $a_1 \leq X \leq a_2$, $b_1 \leq Y \leq b_2$, described by joint probability distribution function $H(x, y)$. Suppose that the function $f(x) \equiv E[Y|X = x]$ is non-decreasing. For any sample of i.i.d. outcomes $(X_1, Y_1), \ldots, (X_n, Y_n)$, if $\hat{f}^n \equiv \hat{f}$ is the estimating function computed by the Monotone Regression Algorithm of Section 2, then for any $\epsilon > 0$,*

$$\Pr\Big\{||\hat{f}^n - f||_2^2 > \epsilon\Big\} \leq 8(n+1)^{\lceil\frac{2(b_2-b_1)}{\epsilon}\rceil} e^{\frac{-\epsilon^2 n}{256(b_2-b_1)^4}},$$

for all $n = 1, 2, \ldots$. In particular, $\hat{f}^n$ converges to f in L_2 norm almost surely.

Remark. Notice, that the theorem does not put any restrictions on the distribution function $H(x, y)$, other than the monotonicity of f. In particular, function f can be discontinuous.

Proof: See the Appendix.

We now turn to a more difficult case of two-dimensional input variable X. Note, that we cannot hope for a distribution free learning using our least squares algorithm. In fact, suppose that the distribution $H(x)$ of the input variable X has the diagonal $\triangle = \{(t, 1-t) : t \in [0, 1]\}$ as a support. Then none of the points $X_1, X_2, \ldots, X_n$, drawn according to $H(x)$, are comparable, and the algorithm will overfit. However, we now prove that if the distribution of the input variable X has a continuous positive density in the domain $[0, 1]$, then our algorithm provides a consistent estimation. If $H(x)$ has a density function $h(x)$, let $h^{X_i}(x_i)$, and $h^{X_i}(x_i|\cdot), i = 1, 2$ denote the marginal and conditional densities of the i-th coordinate.

Theorem 2 *Consider bounded random variables $X \in [0, 1]^2$, $Y \in [0, 1] \subset \Re$, described by joint probability distribution function $H(x, y)$, such that the marginal distribution $H(x)$ of the input variable X has a continuous positive density function $h(x)$ on $[0, 1]^2$. Suppose that the function $f(x) \equiv E[Y|X = x]$ is coordinate-wise non-decreasing on $[0, 1]^2$.*

For any sample of i.i.d. outcomes $(X_1, Y_1), \ldots, (X_n, Y_n)$, let $\hat{f}^n \equiv \hat{f}$ be the estimating function computed by the Monotone Regression Algorithm of Section 2. There exist a constant $\beta > 0$ such that for any $\epsilon > 0$ and for all n

$$\Pr\Big\{||\hat{f}^n - f||_2^2 > \epsilon\Big\} \leq \beta e^{\lceil\frac{4(b_2-b_1)}{\epsilon}\rceil\sqrt{\gamma n} - \frac{\epsilon^2 n}{256(b_2-b_1)^4}}, \tag{8}$$

where

$$\gamma = \sup_{0 \le x_1 \le x_1' \le 1, 0 \le x_2 \le 1} \frac{h^{X_2}(x_2|x_1)}{h^{X_2}(x_2|x_1 \le X_1 \le x_1')}.$$

In particular, $\hat{f}^n$ converges to f in L_2 norm almost surely.

Remark. By assumption, the density function $h(x) = h(x_1, x_2)$ is continuous and positive on $[0,1]^2$. Since $[0,1]^2$ is compact, function $h(x)$ achieves certain positive minimum and maximum values. It follows that γ must be finite. Note that if the coordinates X_1, X_2 of the input variable are independent, then $\gamma = 1$.

Proof: See the Appendix.

The theorem above raises the question to what degree the upper bound (8) is tight? This reduces to the question, what is the divergence rate of the expected VC entropy $N^{\Im}(\epsilon, n)$ corresponding to the family of coordinate-wise monotone functions? As we will see from the proof, this divergence rate is upper bounded by $\Theta(e^{\theta(n^{\frac{1}{2}})})$, when the distribution $H(x,y)$ has a continuous positive density. We now show that the rate $N^{\Im}(\epsilon, n) \approx \Theta(e^{\theta(n^{\frac{1}{2}})})$ is achieved, when the coordinates X_1, X_2 of the input variable X are independent.

Proposition 4 *Suppose that the the coordinates X_1, X_2 of the input variable X are independent. Namely, $H(x) = H^{X_1}(x_1)H^{X_2}(x_1)$. Then the corresponding expected labeling number satisfies*

$$L(m, n, H) \ge 2^{(1-e^{-1})\sqrt{n}}.$$

If the output variable Y satisfies $Y \in \{b_1, b_2\}$, then

$$N^{\Im}(\epsilon, n) \ge 2^{(1-e^{-1})\sqrt{n}}.$$

Proof: See the Appendix.

4 SIMULATIONS RESULTS

In this section, we provide some simulation results involving our monotone regression algorithm from Section 2. We also compare the performance of the algorithm and kernel regression on the same samples of artificially generated data. We use the following model

$$Y = f(X) + \psi,$$

where $0 \leq X \leq 1$, $f(x) = x$, $x \in [0, 1/2]$ and $f(x) = x + 1$, $x \in [1/2, 1]$. In particular, function f is discontinuous. We draw $X_i, i = 1, 2, \ldots, n$ independently and uniformly from $[0, 1]$ and we draw the noise terms independently with normal distribution $N(0, \sigma^2)$. After running Step 1 of the Monotone Regression Algorithm on the values $X_i, Y_i = f(X_i) + \psi_i, i = 1, 2, \ldots, n$ we obtain approximating values $\hat{f}_1, \ldots, \hat{f}_n$. We also compute kernel regression estimates of the function $f(x), x \in [0, 1]$ using the same samples (X_i, Y_i), $i = 1, 2, \ldots, n$ and kernel function $\phi(x, x_0) = e^{-\frac{(x-x_0)^2}{\delta^2}}$. The performance measure

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

is used which approximates the average distance between $f(x)$ and $\hat{f}(x)$. We summarize the results of this experiment in Table 1, corresponding to sample sizes $n = 200$ and performance measure d. The second column lists the values of the performance d obtained by our algorithm. The last three columns are the results of kernel regression estimation using the same data samples and various bandwidths $\delta = .01\Delta, .05\Delta, .1\Delta$, where Δ is a sample variance. We see that our algorithm performs comparably with kernel regression for $\sigma = 1, .5$ and significantly better for smaller values of σ. In particular, the performance is not affected by the discontinuity of the regression function. We also plot the result of running the algorithm on sample size $n = 300$ and variance $\sigma = .5$ on Figure 1. The solid line is the regression function f, the stars are the approximating values $\hat{f}_i, i = 1, 2, \ldots, 300$, the "pluses" represent the actual observations.

$n = 200$	MR	Kernel regression		
σ	Algorithm	$.01\Delta$	$.05\Delta$	$.1\Delta$
1.0000	0.2897	0.3859	0.2360	0.1899
0.5000	0.1488	0.2331	0.1723	0.1727
0.1000	0.0393	0.0740	0.0945	0.1388
0.0500	0.0206	0.0609	0.1061	0.1445
0.0100	0.0065	0.0485	0.0836	0.1295
0.0050	0.0040	0.0560	0.1229	0.1663
0.0010	0.0010	0.0554	0.1178	0.1597
0	0.0000	0.0666	0.1074	0.1581

Table 1.

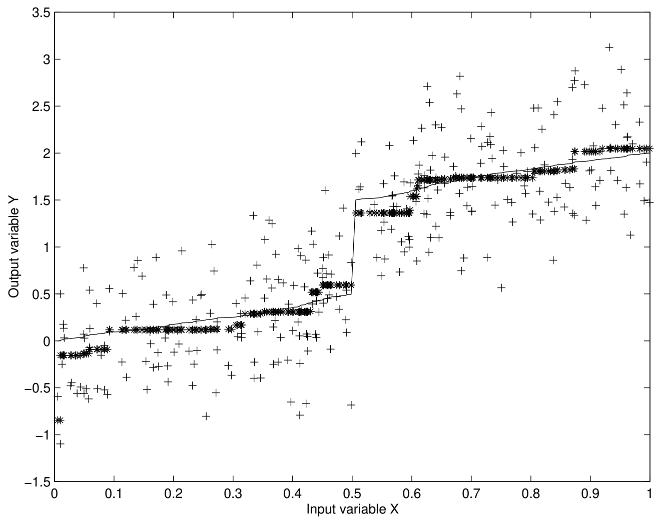

Figure 1: Monotone regression algorithm on a sample size n=300 and noise std=.5

We now consider a two-dimensional regression model

$$Y = \frac{1}{4}X_1 + 2X_2 + \psi,$$

where X_1, X_2 are drawn uniformly and independently from $[0, 1]$, and the noise is distributed as $N(0, \sigma^2)$. We run the algorithm for $\sigma = .5$ on samples $X_i = (X_i^1, X_i^2), Y_i, i = 1, 2, \ldots, n$ for different n. The resulting estimating values $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_n$ are measured against the true values $f(X_i)$, $i = 1, 2, \ldots, n$ using the distance function d. The number of generated constraints (3), which now depends on each particular instance, is also reported.

Sample size n	Performance d	Constraints
20	.4042	98
40	.3385	340
60	.3116	814
80	.2546	1552
100	.2158	2529
120	.2338	3827

Table 2.

5 APPLICATION TO NON-STATIONARY POISSON PROCESSES

In this section we discuss an application of our algorithm to learning an arrival rate of a non-stationary Poisson process. Suppose we are given a non-stationary Poisson arrival process $(N(t), \lambda(t))$, where $N(t)$ denotes the number of arrivals up to time t, and a right-continuous function $\lambda(t)$ denotes a time variable arrival rate. In particular, if $T_1, T_2, \ldots, T_n, \ldots$ denote the arrival times, then

$$\Pr\{T_{k+1} - T_k \geq t | T_k = x\} = e^{-\int_x^{x+t} \lambda(\tau) d\tau}. \tag{9}$$

Non-stationary Poisson processes have many applications in engineering. They are used to model, for example, stream of photons from an optical modulator, call initiations in telephone networks, or customer arrivals in service queues.

We address the question of estimating an arrival rate function $\lambda(t)$ from a sequence of observed arrivals $T_1, T_2, \ldots, T_n$, when the arrival rate function is an unknown non-decreasing right-continuous function. Note from (9), that the density function $f_k(t)$ of the k-th interarrival time $T_{k+1} - T_k$, conditioned on $T_k = x$ is given by

$$f_k(t) = \lambda(x+t) e^{-\int_x^{x+t} \lambda(\tau) d\tau}.$$

Let us formulate a maximum likelihood estimation problem. Given a sequence $T_1, T_2, \ldots, T_n$ of arrival times, we maximize

$$\log \prod_{i=1}^{n} \hat{\lambda}(T_i) e^{-\int_{T_{i-1}}^{T_i} \hat{\lambda}(\tau) d\tau} = \sum_{i=1}^{n} \log \hat{\lambda}(T_i) - \int_0^{T_n} \hat{\lambda}(\tau) d\tau, \tag{10}$$

over the set of all non-decreasing right-continuous functions $\hat{\lambda}(t)$, defined on the domain $[0, T_n]$. Note, that if the values of a function $\hat{\lambda}$ are fixed at points $T_1, T_2, \ldots, T_n$ to be $\hat{\lambda}_1 \leq \hat{\lambda}_2 \leq \cdots \leq \hat{\lambda}_n$, then the maximum of the objective function (10) is obtained by function

$$\hat{\lambda}(t) = \hat{\lambda}_i, \text{ for } t \in (T_{i-1}, T_i],\ i = 1, 2, \ldots, n-1 \hat{\lambda}(t) = 0 \text{ for } t \in [0, T_1], \tag{11}$$

and

$$\int_0^{T_n} \hat{\lambda}(\tau) d\tau = \sum_{i=1}^{n-1} \hat{\lambda}_i (T_{i+1} - T_i).$$

Thus, we reformulate the maximum likelihood problem as the following maximization problem in variables $\hat{\lambda}_i$, $i = 1, 2, \ldots, n$.

$$\text{maximize} \sum_{i=1}^{n} \log \hat{\lambda}_i - \sum_{i=1}^{n-1} \hat{\lambda}_i (T_{i+1} - T_i)$$

subject to

$$\hat{\lambda}_i - \hat{\lambda}_{i+1} \leq 0 \qquad i = 1, 2, \ldots, n-1,$$

$$\hat{\lambda}_1 \geq 0.$$

This optimization problem can be solved in polynomial time, since the objective function is concave, and all constraints are linear (see [7]). Solving this optimization problem we obtain estimates $\hat{\lambda}_i$ of the function $\lambda(t)$ at points T_i. We then use (11) to construct an estimate function $\hat{\lambda}(t)$, defined on the domain $[0, T_n]$.

6 CONCLUSION

We have proposed an algorithm for estimating regression functions with coordinate-wise monotonicity property. The algorithm is based on quadratic optimization. It runs in polynomial time and works even if the regression function is discontinuous. The consistency property of our estimation procedure was proven when the number of observations goes to infinity, for the case of one-dimensional input variables and for the case of two-dimensional input variables, whose distribution has a continuous positive density. Specifically, for two-dimensional input variables we showed that the expected VC entropy grows superpolynomially and subexponentially with the sample size.

7 APPENDIX

Proof of Proposition 3: Introduce $N = 2\lceil \frac{b_2 - b_1}{\epsilon} \rceil$ and $q_k = b_1 + \frac{k}{N}(b_2 - b_1),\ k = 0, 1, 2, \ldots, N$. Consider the set of vectors

$$G =$$

$$\left\{ \left((y_1 - g_1)^2, (y_2 - g_2)^2, \ldots, (y_n - g_n)^2 \right) : g_i \in \{q_0, q_1, \ldots, q_N\}, g_i \leq g_j, \text{ whenever } x_i \prec x_j,\ i, j = 1, 2, \ldots, n \right\}.$$

Then G is a subset of the set (5). Clearly, the cardinality of the set G is equal to $L(N, x_1, x_2, \ldots, x_n)$. To prove the proposition we show that G is an ϵ net of the set (5), for $Q(x, y, \hat{f}) = (y - \hat{f}(x))^2$. Fix $\hat{f} \in \Im$. We now construct a vector which belongs to G and whose distance from $\hat{f}$ is smaller than ϵ. For each $i = 1, 2, \ldots, n$ find k_i such that

$$\hat{f}(x_i) \in [q_{k_i}, q_{k_i+1}). \tag{12}$$

(Assume $k_{N+1} = k_N = b_2$). This is always possible since $\hat{f}(x) \in [b_1, b_2]$, for all $x, \hat{f}$. We have

$$|(y_i - \hat{f}(x_i))^2 - (y_i - q_{k_i})^2| = |\hat{f}(x_i) - q_{k_i}||2y_i - \hat{f}(x_i) - q_{k_i}| \leq \frac{1}{N}2(b_2 - b_1) < \epsilon,$$

where the inequalities follow from (12), the choice of N and from the fact $\hat{f}(x_i), y_i, q_{k_i} \in [b_1, b_2]$. We conclude that the $||\cdot||_\infty$ distance between the vectors $((y_1 - \hat{f}(x_1))^2, \ldots, (y_n - \hat{f}(x_n))^2)$ and $((y_1 - q_{k_1})^2, \ldots, (y_n - q_{k_n})^2)$ is not biger than ϵ. It remains to show that the second vector belongs to G. If $x_i \prec x_j$ then $\hat{f}(x_i) \leq \hat{f}(x_j)$, which implies $q_{k_i} \leq q_{k_j}$.

To prove the lower bound, fix $\epsilon > 0$ and let $N = \lfloor \frac{b_2 - b_1}{\sqrt{2\epsilon}} \rfloor$. Suppose there exists an ϵ-net of the set (5) with cardinality smaller than $L(N, x_1, x_2, \ldots, x_n)$. Construct the set G as above. For each point in the net, let us find point in G which is ϵ close to it (this is possible, since G is a subset of the set (5)). Then we obtain a proper subset of the set G which is a 2ϵ-net. Consider an element of G which is not in the net. We argue that its distance to any other point from G is not smaller than

$$\frac{(b_1 - b_2)^2}{N^2} \geq 2\epsilon$$

and thereby obtain the contradiction to the choice N. Take any two distinct point in G. They must differ by at least one coordinate. Let these coordinates be $(y_i - q_{k_i})^2$ and $(y_i - q_{k'_i})^2$. The difference between these two coordinates is

$$|(y_i - q_{k_i})^2 - (y_i - q_{k'_i})^2| = |q_{k_i} - q_{k'_i}||2y_i - q_{k_i} - q_{k'_i}|$$

But $|q_{k_i} - q_{k'_i}| \geq (b_2 - b_1)/N$. Also since $y_i = b_1$ or $y_i = b_2$, and $q_{k_i}, q_{k'_i} \in [b_1, b_2]$, then $|2y_i - q_{k_i} - q_{k'_i}| \geq (b_2 - b_1)/N$. We conclude that the distance is at least $(b_2 - b_1)^2/N^2$. □

Proof of Theorem 1: Applying Proposition 2, it suffices to show that

$$N(\epsilon, \Im, (x_1, y_1), \ldots, (x_n, y_n)) \leq (n+1)^{\lceil \frac{2(b_2 - b_1)}{\epsilon} \rceil}, \tag{13}$$

where, $\Im$ is the space of non-decreasing functions $\hat{f} : [a_1, a_2] \to [b_1, b_2]$. From Proposition 3

$$N(\epsilon, \Im, (x_1, y_1), \ldots, (x_n, y_n)) \leq L(2\lceil \frac{b_2 - b_1}{\epsilon} \rceil, x_1, \ldots, x_n).$$

Let $N = 2\lceil \frac{b_2 - b_1}{\epsilon} \rceil$. Since $x_1, x_2, \ldots, x_n$ is a fully ordered set, then the labeling number of this set is equal to the number of nondecreasing sequences $z = (z_1, z_2, \ldots, z_n)$ where $z_i \in \{1, 2, \ldots, N\}$. For each sequence z and $k = 1, 2, \ldots, N$ let $i(k)$ be the smallest index i for which $z_i \geq k$d. We put $i(k) = n + 1$ if $z_n < k$. Note, it is not excluded that $i(k) = i(k+1)$, which means no z_i is equal to k. The sequence

$i(k), k = 1, 2, \ldots, N$ uniquely determines z. There are no more than $(n+1)^N$ different sequences $i(k)$. Since $(n+1)^N \leq (n+1)^{\lceil \frac{2(b_2-b_1)}{\epsilon} \rceil}$, (13) is proven. We conclude that $\hat{f}^n$ converges to f in L_2 norm in probability. Since the tail probabilities are exponentially small, then the convergence is almost sure, by Borel-Cantelli lemma. This completes the proof of the theorem $\square$

Proof of Theorem 2: Combining Propositions 2 and 3 it suffices to show that the expected labeling number $L(m, n, H)$ corresponding to the marginal distribution $H(x)$ of the input variable X satisfies

$$L(m, n, H) \leq \beta e^{2m\sqrt{\gamma n}}. \tag{14}$$

We start by reducing the problem to the case of labeling number with two labels.

Lemma 3 *For any sequence $x_i \in \mathcal{D}, i = 1, 2, \ldots, n$*

$$L(m, x_1, \ldots, x_n) \leq L^{m-1}(2, x_1, \ldots, x_n), \quad m = 2, 3, \ldots.$$

Proof: The proof is by induction by m. The inequality is trivial for $m = 2$. Suppose it holds for m and we have points $x_1, x_2, \ldots, x_n \in \mathcal{D}$. Consider any valid labeling

$$\phi : \{x_1, x_2, \ldots, x_n\} \to \{1, 2, \ldots, m+1\}.$$

Consider also the following two labelings

$$g_1(x_i) = \begin{cases} g(x_i), & \text{if } g(x_i) \leq m, \\ m, & \text{if } g(x_i) \leq m+1. \end{cases}$$

$$g_2(x_i) = \begin{cases} g(x_i), & \text{if } g(x_i) \leq m \\ 2, & \text{if } g(x_i) = m+1. \end{cases}$$

Note that both of the labelings are valid (are consistent with the $\prec$ order) and the map $g \to (g_1, g_2)$ is one-to-one. Also g_1 is a valid m labeling and g_2 is a valid 2 labelig. It follows then

$$L(m+1, x_1, \ldots, x_n) \leq L(m, x_1, \ldots, x_n) L(2, x_1, \ldots, x_n).$$

But by induction hypothesis

$$L(m, x_1, \ldots, x_n) \leq (L(2, x_1, \ldots, x_n))^{m-1}.$$

Thus, the bound holds for $L(m+1, x_1, \ldots, x_n)$ as well. $\square$

Using the result above we may now concentrate on establishing the bounds on the expected two-labeling number $L(2,n,H)$. Our objective is to show that for a certain constant β

$$L(2,n,H) \leq \beta e^{2\gamma^{\frac{1}{2}} n^{\frac{1}{2}}}. \tag{15}$$

for all n. Let $\mathcal{D}$ denote the collection of sub-rectangles

$$\Delta = [a_1, a_2] \times [b_1, b_2] \subset [0,1]^2,$$

and let $H_\Delta = H_\Delta(x,y)$ denote the joint distribution of (X,Y), drawn according to H, conditioned on $(X,Y) \in \Delta$. Denote by h_Δ the density of H_Δ.

Lemma 4 *There holds*

$$\sup_{\Delta\in\mathcal{D}} L(2,n,H_\Delta) \leq \sup_{\Delta\in\mathcal{D}} L(2,n-1,H_\Delta) + \frac{\gamma}{n}\sum_{k=0}^{n-1} \sup_{\Delta\in\mathcal{D}} L(2,k,H_\Delta).$$

Proof: We fix $\Delta = [a_1,a_2]\times[b_1,b_2] \in \mathcal{D}$ and choose $(X_1,Y_1),\ldots,(X_n,Y_n)$ according to H_Δ. Select an index $i = i_{\min}$ such that $X_i \geq X_{i_{\min}}$ for all i. Note that if we label the point $(X_{\min}, Y_{\min})$ with 1, then the remaining $n-1$ points can be labeled independently of $(X_{\min}, Y_{\min})$ (since none of them has X component smaller than $X_{\min}$). On the other hand, if we label $(X_{\min}, Y_{\min})$ with 2, then all the points with $Y_k \geq Y_{\min}$ (thus satisfying $(X_{\min}, Y_{\min}) \prec (X_k, Y_k)$) are constrained to be labeled 2, and all the remaining points can be labeled independently from $(X_{\min}, Y_{\min})$. We conclude that

$$L(2,n,H_\Delta) = L(2,n-1,H_\Delta) + E[L(2,(X_{i_1},Y_{i_1}),\ldots,(X_{i_k},Y_{i_k})], \tag{16}$$

where $Y_{i_1},\ldots,Y_{i_k}$ are the values from the set $\{Y_1,\ldots,Y_n\}$ satisfying $Y_{i_1} \leq Y_1,\ldots,Y_{i_k} \leq Y_1$. Let us fix $(x,y) \in \Delta$. We now estimate the second summand in the recursion (16) conditioned on the event B that $X_{\min} = x, Y_{\min} = y$ and exactly k out of remaining $n-1$ points fall into the rectangle $[x,a_2]\times[b_1,y]$ (and the remaining $n-1-k$ fall into $[x,a_2]\times[y,b_2]$). Note that conditioned on B, we have k points in the rectangle $[x,a_2]\times[b_1,y]$ distributed according to $H_{[x,a_2]\times[b_1,y]}$. Therefore, conditioned on B the second summand in (16) satisfies

$$E[L(2,(X_{i_1},Y_{i_1}),\ldots,(X_{i_k},Y_{i_k}))|B] = L(2,k,H_{[x,a_2]\times[b_1,y]}) \leq \sup_{\Delta\in\mathcal{D}} L(2,k,H_\Delta). \tag{17}$$

Combining with (16) and (17) we obtain

$$L(2,n,H_\Delta) \leq L(2,n-1,H_\Delta) + \sum_{k=0}^{n-1} \left(\begin{array}{c} n-1 \\ k \end{array} \right) \sup_{\Delta\in\mathcal{D}} L(2,k,H_\Delta) \int_{a_1}^{a_2}\int_{b_1}^{b_2} \Pr\{B\}dxdy \tag{18}$$

It is straightforward to compute that the probability of the event B is equal to

$$\Pr\{B\} = \begin{pmatrix} n-1 \\ k \end{pmatrix} \Pr^k\{X > x, Y < y\} \Pr^{n-1-k}\{X > x, Y > y\} h_\Delta(x,y) =$$

$$\begin{pmatrix} n-1 \\ k \end{pmatrix} \Pr^k\{Y < y | X > x\}(1 - \Pr\{Y < y|X > x\})^{n-k-1} \Pr^{n-1}\{X > x\} h_\Delta(x,y), \tag{19}$$

where (X,Y) is a random vector chosen according to the distribution H_Δ with density function h_Δ.

We have

$$h_\Delta(x,y) = h^Y_\Delta(y|X = x) h^X_\Delta(x) = h^Y_\Delta(y|X > x) h^X_\Delta(x) \frac{h^Y_\Delta(y|X = x)}{h^Y_\Delta(y|X > x)}.$$

Note that

$$\frac{h^Y_\Delta(y|X = x)}{h^Y_\Delta(y|X > x)} = \frac{h^Y(y|X = x)}{h^Y(y|x \le X \le a_2)} \le \gamma,$$

by the choice of γ. Therefore

$$h_\Delta(x,y) \le \gamma h^Y_\Delta(y|X > x) h^X_\Delta(x). \tag{20}$$

Observe

$$h^Y_\Delta(y|X > x) = \frac{d\Pr\{Y < y|X > x\}}{dy}.$$

We now evaluate $\int_{a_1}^{a_2}\int_{b_1}^{b_2} \Pr\{B\} dxdy$ using (19). We first integrate by y. From the previous equality, applying integration by parts and using $\Pr\{Y < b_1|X > x\} = 0$, $\Pr\{Y < b_2|X > x\} = 1$ we obtain

$$\int_{b_1}^{b_2} \Pr^k\{Y < y|X > x\}(1 - \Pr\{Y < y|X > x\})^{n-k-1} h^Y_\Delta(y|X > x) =$$

$$\int_{b_1}^{b_2} \Pr^k\{Y < y|X > x\}(1 - \Pr\{Y < y|X > x\})^{n-k-1} d\Pr\{Y < y|X > x\} =$$

$$\frac{1}{k+1}\int_{b_1}^{b_2} (1 - \Pr\{Y < y|X > x\})^{n-k-1} d\Pr^{k+1}\{Y < y|X > x\} = \cdots$$

$$= \frac{(n-1-k)(n-2-k)\cdots 2\cdot 1}{(k+1)(k+2)\cdots(n-1)} = \frac{1}{n\begin{pmatrix} n-1 \\ k \end{pmatrix}}.$$

Also

$$\int_{b_1}^{b_2} \Pr^{n-1}\{X > x\} h^X_\Delta(x) dx = \frac{1}{n}.$$

Combining the previous two equalities with (18),(19) and (20), we obtain

$$L(2,n,H_\Delta) \le L(2,n-1,H_\Delta) + \sum_{k=0}^{n-1} \sup_{\Delta\in\mathcal{D}} L(2,k,H_\Delta)\frac{1}{n}\gamma.$$

It follows

$$\sup_{\Delta\in\mathcal{D}} L(2,n,H_\Delta) \le \sup_{\Delta\in\mathcal{D}} L(2,n-1,H_\Delta) + \frac{\gamma}{n}\sum_{k=0}^{n-1}\sup_{\Delta\in\mathcal{D}} L(2,k,H_\Delta).$$

This completes the proof. $\square$

We use the recursion in Lemma 4 to deduce the rate of divergence of $\sup_{\Delta\in\mathcal{D}} L(2,n,H_\Delta)$ as n diverges to infinity. We argue that the rate of divergence is dominated by $e^{2\sqrt{\gamma n}}$, which has a subexponential growth rate.

Lemma 5 *For all sufficiently large n*

$$e^{2\gamma^{\frac{1}{2}}n^{\frac{1}{2}}} \ge e^{2\gamma^{\frac{1}{2}}(n-1)^{\frac{1}{2}}} + \frac{\gamma}{n}\sum_{k=0}^{n-1} e^{2\gamma^{\frac{1}{2}}k^{\frac{1}{2}}}. \tag{21}$$

Proof: Introduce

$$\alpha(n) = e^{2\gamma^{\frac{1}{2}}n^{\frac{1}{2}}} - e^{2\gamma^{\frac{1}{2}}(n-1)^{\frac{1}{2}}} - \frac{\gamma}{n}\sum_{k=0}^{n-1} e^{2\gamma^{\frac{1}{2}}k^{\frac{1}{2}}}.$$

We need to show that $\alpha(n)\ge 0$ for sufficiently large n. We have

$$n\alpha(n) = ne^{2\gamma^{\frac{1}{2}}n^{\frac{1}{2}}} - ne^{2\gamma^{\frac{1}{2}}(n-1)^{\frac{1}{2}}} - \gamma\sum_{k=0}^{n-1} e^{2\gamma^{\frac{1}{2}}k^{\frac{1}{2}}}.$$

Similarly

$$(n+1)\alpha(n+1) = (n+1)e^{2\gamma^{\frac{1}{2}}(n+1)^{\frac{1}{2}}} - (n+1)e^{2\gamma^{\frac{1}{2}}n^{\frac{1}{2}}} - \gamma\sum_{k=0}^{n} e^{2\gamma^{\frac{1}{2}}k^{\frac{1}{2}}}.$$

Subtracting, we obtain

$$(n+1)\alpha(n+1) - n\alpha(n) = (n+1)e^{2\gamma^{\frac{1}{2}}(n+1)^{\frac{1}{2}}} - (2n+1+\gamma)e^{2\gamma^{\frac{1}{2}}n^{\frac{1}{2}}} + ne^{2\gamma^{\frac{1}{2}}(n-1)^{\frac{1}{2}}}. \tag{22}$$

Now, let us consider the function $e^{2\gamma^{\frac{1}{2}}x^{\frac{1}{2}}}$ and its third order Taylor expansion around $x=n$.

$$e^{2\gamma^{\frac{1}{2}}(n+1)^{\frac{1}{2}}} = e^{2\gamma^{\frac{1}{2}}n^{\frac{1}{2}}}\Big[1 + \frac{\gamma^{\frac{1}{2}}}{n^{\frac{1}{2}}}(n+1-n) + \frac{1}{2}\frac{\gamma}{n}(n+1-n)^2 - \frac{1}{4}\frac{\gamma^{\frac{1}{2}}}{n^{\frac{3}{2}}}(n+1-n)^2 +$$

$$\frac{1}{6}(\frac{\gamma^{\frac{3}{2}}}{n^{\frac{3}{2}}} - \frac{3\gamma}{2n^2} + \frac{3\gamma^{\frac{1}{2}}}{4n^{\frac{5}{2}}})(n+1-n)^3 + O(\frac{1}{n^2})\Big].$$

Similarly,

$$e^{2\gamma^{1}(n-1)^{\frac{1}{2}}} = e^{2\gamma^{\frac{1}{2}}n^{\frac{1}{2}}}\Big[1 + \frac{\gamma^{\frac{1}{2}}}{n^{\frac{1}{2}}}(n-1-n) + \frac{1}{2}\frac{\gamma}{n}(n+1-n)^2 - \frac{1}{4}\frac{\gamma^{\frac{1}{2}}}{n^{\frac{3}{2}}}(n+1-n)^2 +$$

$$\frac{1}{6}(\frac{\gamma^{\frac{3}{2}}}{n^{\frac{3}{2}}} - \frac{3\gamma}{2n^2} + \frac{3\gamma^{\frac{1}{2}}}{4n^{\frac{5}{2}}})(n-1-n)^3 + O(\frac{1}{n^2})\Big].$$

Applying this to (22), and simplifying we obtain

$$(n+1)\alpha(n+1)-n\alpha(n) = e^{2\gamma^{\frac{1}{2}}n^{\frac{1}{2}}}\Big[-\gamma+\frac{\gamma^{\frac{1}{2}}}{n^{\frac{1}{2}}}+\gamma+\frac{1}{2}\frac{\gamma}{n}-\frac{1}{2}\frac{\gamma^{\frac{1}{2}}}{n^{\frac{1}{2}}}-\frac{1}{4}\frac{\gamma^{\frac{1}{2}}}{n^{\frac{3}{2}}}+(n+1-n)\frac{1}{6}(\frac{\gamma^{\frac{3}{2}}}{n^{\frac{3}{2}}}-\frac{3\gamma}{n^2}+\frac{3\gamma^{\frac{1}{2}}}{4n^{\frac{5}{2}}})+O(\frac{1}{n})\Big] =$$

$$= e^{2\gamma^{\frac{1}{2}}n^{\frac{1}{2}}}\Big[\frac{1}{2}\frac{\gamma^{\frac{1}{2}}}{n^{\frac{1}{2}}}+O(\frac{1}{n})\Big] \geq e^{2\gamma^{\frac{1}{2}}n^{\frac{1}{2}}}\frac{1}{4}\frac{\gamma^{\frac{1}{2}}}{n^{\frac{1}{2}}},$$

for sufficiently large n. Since the sum

$$\sum_{n=1}^{\infty} e^{2\gamma^{\frac{1}{2}}n^{\frac{1}{2}}}\frac{1}{4}\frac{\gamma^{\frac{1}{2}}}{n^{\frac{1}{2}}}$$

diverges to infinity, then

$$n\alpha(n)-\alpha(1)=\sum_{m=1}^{n-1}(m+1)\alpha(m+1)-m\alpha(m)$$

diverges to infinity as well. In particular, $\alpha(n)$ is positive for all sufficiently large n. $\square$

Our final step in proving Theorem 2 is to use Lemma 5 to show that there exists some constant $\beta>0$ for which $\sup_{\Delta\in\mathcal{D}} L(2,n,H_\Delta)$ is upper bounded by $\beta e^{2\gamma^{\frac{1}{2}}n^{\frac{1}{2}}}$ for all n. Let us find n_0 starting from which (21) holds. We now show by induction that

$$\sup_{\Delta\in\mathcal{D}} L(2,n,H_\Delta)\leq 2^{n_0}e^{2\gamma^{\frac{1}{2}}n^{\frac{1}{2}}},$$

for all $n=1,2,\ldots$. Since $L(2,n,H_\Delta)\leq 2^n$ (the labeling number of n points cannot exceed 2^n) then the inequality holds for all $n\leq n_0$. Suppose it holds for all $n\leq m-1$ where $m>n_0$. Then applying Lemma 4

$$\sup_{\Delta\in\mathcal{D}} L(2,m,H_\Delta)\leq 2^{n_0}\Big(e^{2\gamma^{\frac{1}{2}}(m-1)^{\frac{1}{2}}}+\frac{\gamma}{m}\sum_{k=0}^{m-1}e^{2\gamma^{\frac{1}{2}}k^{\frac{1}{2}}}\Big)\leq 2^{n_0}e^{2\gamma^{\frac{1}{2}}m^{\frac{1}{2}}},$$

where the last inequality follows from Lemma 5 and $m>n_0$. This proves the induction step.

We conclude that for the constant $\beta=2^{n_0}$ there holds

$$\sup_{\Delta\in\mathcal{D}} L(2,n,H_\Delta)\leq \beta e^{2\gamma^{\frac{1}{2}}n^{\frac{1}{2}}}.$$

In particular, let us take the rectangle C to be the whole domain $\mathcal{D}$. Then (8) is satisfied. We conclude that the bound (15) holds, namely $\hat{f}^n$ converges to f in probability, exponentially fast. The almost sure convergence follows from Borel-Cantelli lemma. $\square$

Proof of the Proposition 4: Without the loss of generality, we may assume that the input distribution $H(x)$ is uniform. We have $L(m,n,H)\geq L(2,n,H)$ and it suffices to prove the lower bound for the case of two labels $1,2$ only.

Let $N = \lceil \sqrt{n} \rceil$ and let $p_k^1 = a_1^1 + \frac{k}{N}(a_2^1 - a_1^1), p_k^2 = a_1^2 + \frac{k}{N}(a_2^2 - a_1^2), k = 0, 2, \ldots, N-1$. Then we have partitioned the rectangle $\mathcal{D}$ into N^2 equal size rectangles $P_{kj} \equiv [p_k^1, p_{k+1}^1] \times [p_j^2, p_{j+1}^2]$. Let us consider the rectangles forming anti-diagonal $P_{N-1,0}, P_{N-2,1}, \ldots, P_{0,N-1}$. Given random input points $X_1, X_2, \ldots, X_n$, for each $k = 0, 1, \ldots, N-1$ select one point out of $X_1, X_2, \ldots, X_n$ that falls into the rectangle $P_{k,N-1-k}$, if there is any. Let $\Delta = \Delta(X_1, X_2, \ldots, X_n)$ denote number of points selected. In particular, $\Delta \leq N$. Notice, that these points are incomparable with respect to $\prec$ order, and therefore can be labeled with $1, 2$ arbitrarily. Any of such labelings can be extended to a labeling over all points $X_1, X_2, \ldots, X_n$. We conclude

$$L(2, n, H) \geq E[2^{\Delta}]$$

By Jensen's inequality $E[2^{\Delta}] \geq 2^{E[\Delta]}$. We now show that $E[\Delta] \geq \sqrt{n}(1 - e^{-1})$. Since all the rectangles have equal size, the probability of a point falling into a given rectangle $P_{k,N-1-k}$ is $1/N^2$. Therefore the probability that at least one point falls into $P_{k,N-1-k}$ is $1 - (1 - 1/N^2)^{N^2}$. It follows

$$E[\Delta] \geq N[1 - (1 - 1/N^2)^{N^2}]$$

Since $(1 - 1/N^2)^{N^2} \leq e^{-1}$, then

$$E[\Delta] \geq N(1 - e^{-1}) \geq \sqrt{n}(1 - e^{-1}).$$

This proves the first part of the proposition. The second part follows immediately from Proposition 3. $\square$

Acknowledgments. Many thanks to Don Coppersmith for suggesting the proof of the labeling problem and many helpful discussions.

References

[1] D. Bertsimas, D. Gamarnik J. Tsitsiklis. Estimation of Time-Varying Parameters in Statistical Models; an Optimization Approach. *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, 314-324, 1997.

[2] J. Hutchinson, A. Lo and T. Poggio. A Nonparametric Approach to Pricing and Hedging Derivative Securities Via Learning Networks. Technical Report, Artificial Intelligence Laboratory and Center for Biological and Computational Learning, MIT, 1994.

[3] M. Kearns and R. Schapire. Efficient distribution-free learning of probabilistic concepts. *31st Annual Symposium on Foundations of Computer Science*, 382-391, 1990.

[4] 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.

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

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

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

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