Local Minimax Learning of Approximately Polynomial Functions

Lee Jones and Konstantin Rybnikov
University of Massachusetts at Lowell

1 Introduction

Suppose we have a number of noisy measurements of an unknown real-valued function f near a point of interest $\mathbf{x}_0 \in \mathbb{R}^d$. Suppose also that nothing can be assumed about the noise distribution, except for zero mean and bounded covariance matrix. We want to estimate f at $\mathbf{x}_0$ using a general linear parametric family $f(\mathbf{x}; \mathbf{a}) = a_0 h_0(\mathbf{x}) + \ldots + a_q h_q(\mathbf{x})$, where $\mathbf{a} \in \mathbb{R}^q$ and h_i's are bounded functions on a neighborhood B of $\mathbf{x}_0$, which contains all points of measurement. Typically, B is a Euclidean ball or cube in $\mathbb{R}^d$ (more generally, a ball in an l_p-norm). In the case when the h_i's are polynomial functions in $(x_1, \ldots, x_d) = \mathbf{x}$ the model is called locally-polynomial. In particular, if the h_i's form a basis of the linear space of polynomials of degree at most two, the model is called locally-quadratic (if the degree is at most three, the model is locally-cubic, etc.). More generally, h_i are picked from some linear space of functions $\mathcal{H}$, which must include at least all affine functions on $\mathbb{R}^d$. Often, information about the behavior of function f on B is available, which is referred to as *context,* such as, e.g., that f takes values in a known interval, or that it satisfies a Lipschitz condition, etc. Given a loss function and a linear space of functions $\mathcal{H}$, the idea of local minimax learning is in choosing for each point of interest $\mathbf{x}_0$ a parameter vector $\mathbf{a}$ that minimizes the maximal possible loss over all $\mathbf{a} \in \mathbb{R}^q$. *The bounds and algorithms are not based on asymptotics or Bayesian assumptions and are truly local for each query, not depending on cross validating estimates at other queries to optimize modeling.* The theory of local minimax estimation with context for locally-polynomial models and approximately locally polynomial models has been recently initiated by Jones (2006) and the focus of our paper is on a subclass of problems studied by Jones, which reduce to real algebraic geometry and finite-dimensional optimization of linear functions over compact domains. See Jones (2006) for detailed treatment.

Denote by $\mathcal{P}_{r,d}$ the space of real polynomial of degree at most r in d indeterminates. In the case of $\mathcal{H} = \mathcal{P}_{1,d}$ and a given bound on the change of f on $B = \{\mathbf{x} \in \mathbb{R}^d \mid |\mathbf{x}| \leq 1\}$, the solution for the squared error loss function is in the form of ridge regression, where the ridge parameter is identified; hence, a minimax justification for ridge regression is given, together with explicit best error bounds. The analysis of polynomial models of degree above 1 leads to interesting and difficult questions in real algebraic geometry and non-linear optimization. We show that in the

Dagstuhl Seminar Proceedings 06201
Combinatorial and Algorithmic Foundations of Pattern and Association Discovery
http://drops.dagstuhl.de/opus/volltexte/2007/891

case when f is a probability function, the optimal in the minimax sense estimator is effectively computable with any given precision, thanks to Tarski's quantifier elimination principle.

2 General Inverse Problem

The methods described in this paper can be applied to a much larger class of problems than those mentioned in the introduction. That is why we give here a more general formulation of the problem under investigation. Denote by $[n]$ the set $\{1, \ldots, n\}$. Suppose we are given real values Y_j for $j \in [k]$,

$$Y_j = \int\limits_{\mathbb{R}^d} f \, d\mu_j + N_j$$

- where μ_j's are known measures,
- f is an unknown $\mathbb{R}$-valued function on $\mathbb{R}^d$, the value of which is to be estimated at $\mathbf{x}_0$,
- N_j's are random variables with zero mean and covariance matrix $\mathbf{N}$, bounded from above in the semi-definite order ($\mathbf{M} \succeq \mathbf{M}'$ iff $\mathbf{M} - \mathbf{M}'$ is positive semi-definite) by a known positive definite matrix $\mathbf{S}$.

The problem of estimation of the value of f at $\mathbf{x}_0$ is called the General Inverse Problem. In practice, one normally has to assume that f belongs to some rather narrow subspace of the space of all $\mathbb{R}$-valued functions on $\mathbb{R}^d$ and that for each $j \in [k]$

$$\int\limits_{\mathbb{R}^d} f(\mathbf{t}) d\mu_j = \int\limits_{\mathbb{R}^d} \theta_j(\mathbf{t}) f(\mathbf{t}) d\mathbf{t} + \int\limits_{\mathbb{R}^d} \sum_{\mathbf{p} \in I_j} c_{\mathbf{p}} \delta(\mathbf{t} - \mathbf{p}) f(\mathbf{t}) \, d\mathbf{t}$$

where θ_j is a known piecewise-analytic "kernel" function and $|I_j| < \infty$. Furthermore, many problems in applications reduce to an even simpler form of the inverse problem where for each $j \in [k]$

$$Y_j = \int \delta(\mathbf{t} - \mathbf{x}_j) f(\mathbf{t}) d\mathbf{t} = f(\mathbf{x}_j) + N_j.$$

We call this form of the inverse problem the Standard Inverse Problem.

3 Affine Estimators for Standard Inverse Problem

Let $\mathcal{H}$ be a q-dimensional linear space of functions which are bounded on a neighborhood B of $\mathbf{x}_0$. Typically, B is a ball in $(\mathbb{R}^d, l_p)$, where $p \in [1, \infty]$, centered at $\mathbf{x}_0$. Assume that the unknown function f belongs to $\mathcal{H}$, i.e.

$$f(\mathbf{x}; \mathbf{a}) = a_1 h_1(\mathbf{x}) + \ldots + a_q h_q(\mathbf{x})$$

where $\{h_i\}_{i=1}^q$ form a basis of $\mathcal{H}$ and $(a_1, \ldots, a_q) = \mathbf{a} \in \mathbb{R}^q$. We seek to estimate $f(\mathbf{x}_0)$ in the form of $F(\mathbf{w}) = w_0 + \mathbf{w} \cdot \mathbf{Y} = w_0 + \sum_1^k w_j Y_j$. That is, we are looking for an estimator in the space $\mathrm{Aff}(\mathbb{R}^k)$ of affine functions of the measurements Y_j, $j \in [k]$. Since we do not know anything about the distribution of the noise, we need to use some *contextual restrictions* on f, that is the information that is known from the application at hand. Fix B. The following types of context are often encountered in engineering and sciences.

1. f takes values in a known interval (e.g. probability function). Denote the class of functions f such that $f(B) \subset [a, b] \subset \mathbb{R}$ by $\mathcal{C}$-Range$([a, b])$.

2. f does not change between any two points in B by more than a fixed amount. Denote the class of functions f such that $|f(\mathbf{x}) - f(\mathbf{x}')| \leq c$ for any $\mathbf{x}, \mathbf{x}' \in B$ by $\mathcal{C}$-Change(c).

3. f satisfies a Lipschitz condition. Denote the class of functions satisgying the Lipschitz condition with constant C on B by $\mathcal{C}$-Lip(C).

4. $f \in \mathcal{D}^s(\mathbb{R}^d)$ and $f^{(s)}$ satisfies one of the conditions in 1-3). For example, if $f''' \in \mathcal{C}$-Lip(C), we will write $f'' \in \mathcal{C}$-Lip$^2(C)$.

Furthermore, f may satisfy a number of conditions of the above types, e.g. $f \in \mathcal{C}$-Range$([a, b]) \cap \mathcal{C}$-Lip$^3(C)$ stands for all functions in $\mathcal{D}^3(B)$ whose values lie in $[a, b]$ and whose third derivative is Lioshitz with constant C. We will use the same abbreviation to refer to a function class and to the condition that defines this class.

To compare estimators and their errors we need to agree on the way the error is measured. Such a measure is known as a loss function in game theory. By a loss function $L(\mathbf{x}; F, h)$ we mean any non-negative function $L : B \times \mathbb{R} \times \mathcal{H} \to \mathbb{R}_{\geq 0}$. Most often one is concerned with the value of L for $\mathbf{x} = \mathbf{x}_0$; we write $L(F, h)$ for $L(\mathbf{x}_0; F, h)$. The most popular measure in statistics is $L(\mathbf{x}; F, h) = (h(\mathbf{x}) - F)^2$, which is known as the square error. Many other choices are often of type $L(\mathbf{x}; F, h) = |h(\mathbf{x}) - F|^p$, where $p \geq 1$.

Definition 1 *(*Principle of Minimax Learning*) Let $\mathbf{x}_0 \in \operatorname{int} B \subset \mathbb{R}^d$, where B is some body. Let P be a continuous probability distribution on B. Suppose $\mathcal{H}$ is a q-dimensional subspace of $\mathbb{R}^{\mathbb{R}^d}$ that contains* $\mathrm{Aff}(\mathbb{R}^d)$*, L is a loss function, $\mathcal{C}$ is the context, $\mathbf{x}_1, \ldots, \mathbf{x}_k \in B$ are observation points, and $\mathbf{Y} \in \mathbb{R}^k$ is the vector of observations. The principle of local minimax learning states that an* optimal affine estimator $F(\mathbf{w}; \mathbf{Y}) = w_0 + \mathbf{w} \cdot \mathbf{Y} = w_0 + \sum_1^k w_j Y_j$ *for $f \in \mathcal{H}$ at $\mathbf{x}_0$ must be chosen according to*

$$F = \arg\min_F \max_{h \in \mathcal{H} \cap \mathcal{C}} \mathbf{E}_{\mathsf{P}}\{L(\mathbf{x}_0; F, h) \mid \mathbf{x}_1, \ldots, \mathbf{x}_k\}$$

We will use $\mathbf{w}$ for the optimal value of the parameter $\mathbf{v}$ in $F(\mathbf{v}, \mathbf{Y}) = v_0 + \sum_1^k v_j Y_j$ Often we cannot assume that the unknown function f belongs to $\mathcal{H}$, but can assume that f is within some known or conjectured $\varepsilon(\mathbf{x})$ of a member of $\mathcal{H}$. In this case we will say that f is approximately of class

$\mathcal{H}$. For example, when $\mathcal{H} = \mathcal{P}_{1,d}$ we call f approximately affine, etc. Let $\varepsilon : B \to \mathbb{R}_+$. Denote the class of functions there are within $\varepsilon(\mathbf{x})$ of some element of $\mathcal{H}$ by $\mathcal{H}(\varepsilon)$. Then the Principle of Minimal Learning states that an *optimal affine estimator* $F = w_0 + \mathbf{w} \cdot \mathbf{Y} = w_0 + \sum_1^k w_j Y_j$ for $f \in \mathcal{H}(\varepsilon)$ at $\mathbf{x}_0$ must be chosen according to

$$F = \arg\min_F \max_{h \in \mathcal{H}(\varepsilon) \cap \mathcal{C}} \mathbf{E}_{\mathsf{P}}\{L(F,h) \mid \mathbf{x}_1, \ldots, \mathbf{x}_k\}$$

4 Optimal Affine Estimators for Approximately Affine Functions

In this section we will use the following setup. First, without loss of generality assume that $\mathbf{x}_0 = \mathbf{0}$. The neighborhood B of $\mathbf{x}_0 = \mathbf{0}$ that contains the observation points $\mathbf{x}_j$ is the Euclidean ball of radius ρ centered at $\mathbf{0}$. Denote by J the vector of all ones of length k, and let X_0 be the matrix whose rows are the coordinates of $\mathbf{x}_j$'s. Set $\mathbf{X} = [\mathbf{J}|\mathbf{X}_0]$. Let $\mathcal{C}$ denote the class of functions that satisfy the condition that $|g(\mathbf{x}) - g(\mathbf{x}')| \leq 2M$ for any $\mathbf{x}, \mathbf{x}' \in B$. Let $L(\mathbf{x}; G, g) = (G - g(\mathbf{x}))^2$, the squared error loss-function. The following theorem gives an estimate on the error of the optimal affine estimator in the case when it happens to have zero constant term, i.e. when it is linear. Such estimators are called *smoothers* in statistics. Alternatively, this result can be considered as a bound on the error of a linear estimator.

Theorem 2 *Suppose f is within $\varepsilon|\mathbf{x}|^2$ from some function in $\mathcal{H} = \mathcal{P}_{1,d}$. Let $F = w_0 + \mathbf{w} \cdot \mathbf{Y}$ be an optimal affine estimator of f. If $w_0 = 0$, then $L(F, f)$ is bounded from above:*

$$L(F, f) \leq \min_{\mathbf{v} \in \mathbb{R}^k} \{\mathbf{v}^t \mathbf{S} \mathbf{v} + R(\mathbf{v})\},$$

with

$$R(\mathbf{v}) = R(v_1, \ldots v_k) = \begin{cases} \left((\frac{M}{\rho} + \varepsilon\rho) |\sum_1^k v_j \mathbf{x}_j| + \varepsilon \sum_1^k |v_j| \, |\mathbf{x}_j|^2 \right)^2 & \text{if } \varepsilon \leq M\rho^{-2}, \\ \left(\frac{1}{2}(\varepsilon M)^{1/2} |\sum_1^k v_j \mathbf{x}_j| + \varepsilon \sum_1^k |v_j| \, |\mathbf{x}_j|^2 \right)^2 & \text{if } \varepsilon > M\rho^{-2}. \end{cases} \tag{4.1}$$

where $\mathbf{v}$ must satisfy $\sum_{i<k} v_i = v_k$.

Proof. The proof is given in Jones (2006), see Theorem II in there. ■

This theorem gives a robust extension of penalized local linear regression when nonlinearity is present. Note that the above bound does not hold when the optimal estimator has a non-zero affine term. The following theorem deals with the case where f is affine. It is convenient to set $\lambda = (\frac{\rho}{M})^2$

Theorem 3 *Suppose $f \in \mathcal{H} = \mathcal{P}_{1,d}$. Let $F = w_0 + \mathbf{w} \cdot \mathbf{Y}$ be an optimal affine estimator of f. Then $w_0 = 0$ and*

$$\mathbf{w} = \arg\min_{\mathbf{v} \in \mathbb{R}^k} \left\{ \mathbf{v}^t \mathbf{S} \mathbf{v} + \frac{1}{\lambda} |\sum_1^k v_j \mathbf{x}_j|^2 \right\}$$

Furthermore, if we denote by $\mathbf{Q}$ *the matrix* $(\lambda^{-1}\mathbf{X}\mathbf{X}^t + \mathbf{S})^{-1}$ *and by* $\Sigma_{\mathbf{Q}}$ *the sum of its elements, then*

$$\mathbf{w} = \frac{1}{\mathbf{J}^t\mathbf{Q}\mathbf{J}}\mathbf{Q}\mathbf{J},$$

and the error of the optimal estimator is bounded from above by

$$\frac{1}{|\Sigma_{\mathbf{Q}}|} - \frac{1}{\lambda}$$

Proof. The proof is given in Jones (2006), see Theorem II in there. ■

In particular, this theorem presents the correct choice of penalty parameter and new error bounds for what is known in statistics as Ridge Regression by a Gradient Regularization. In the language of the latter method, $F(\mathbf{w})$ is the constant term in gradient regularization and λ is the ridge parameter, also known as penalty. Thus, the above theorem gives a rigorous minimax justification of Ridge Regression by a Gradient Regularization.

5 Optimal Affine Estimators for Approximately Polynomial Functions

In this section we will use the following setup. As befire assume that $\mathbf{x}_0 = \mathbf{0}$. We seek to estimate the value of an unknwon function f at $\mathbf{0}$ using noisy observations of f at points $\mathbf{x}_j$, where $j \in [k]$, which all lie in B, a neighborhood of $\mathbf{x}_0$. The neighborhood $B \subset \mathbb{R}^d$ of $\mathbf{x}_0 = \mathbf{0}$ is a bounded body. Let $\mathcal{H}$ be a q-dimensional subspace of the space of all polynomials in d variables. It is known that there is a function $f_{\mathcal{H}} \in \mathcal{H}$, $\varepsilon \geq 0$ amd $r \geq 1$ such that for all $\mathbf{x} \in B$ we have $|f(\mathbf{x}) - f_{\mathcal{H}}(\mathbf{x})| \leq \varepsilon|\mathbf{x}|^r$. Let us see what has to be done in order to find an optimal affine estimator, that is to find an element of the following subset of $\mathrm{Aff}(R^d)$

$$\arg\min_{F} \max_{h \in \mathcal{H}(\varepsilon|\mathbf{x}|^r) \cap \mathcal{C}} \mathbf{E}_{\mathbf{P}}\{L(\mathbf{x}_0; F, h) \mid \mathbf{x}_1, \ldots, \mathbf{x}_k\}$$

First, we have to characterize the set of functions $\mathcal{H}(\varepsilon|\mathbf{x}|^r) \cap \mathcal{C}$.

Theorem 4 *Let* $\{\mathcal{C}_i\}_{i\in\mathbb{I}}$*, with* $\mathbb{I}$ *an arbitrary index set, be a collection of classes of functions, where each* $\mathcal{C}_i$ *is of one of the four types described in Section 3. Then* $\bigcap\limits_{i\in\mathbb{I}} \mathcal{C}_i$ *is convex.*

Proof. Suppose f_1 and f_2 are both in $\mathcal{C}$-Range($[a, b]$]). Since the addition of functions and the multiplication of a function by a constant are defined pointwise, and $[a, b]$ is convex, $\mathcal{C}$-Range($[a, b]$) is convex. Similarly, it is easy to see that $\mathcal{C}$-Change(c) and $\mathcal{C}$-Lip(C) are convex. Bt linearlity of differentiation the same is true for derived classes $\mathcal{C}$-Range^s($[a, b]$), $\mathcal{C}$-Change^s(c), $\mathcal{C}$-Lip^s(C). The intersection of any family of convex sets is convex. ■

Lemma 5 *For any* $\varepsilon(\mathbf{x}) \geq 0$ *and any* $\mathcal{H}$ *the set* $\mathcal{H}(\varepsilon(\mathbf{x}))$ *is convex.*

Proof. The proof is by direct check. ■

Thus, $\mathcal{H}(\varepsilon|\mathbf{x}|^r)\cap\mathcal{C}$ is convex. In applications most always B is a body with piecewise-algebraic boundary and $\varepsilon(\mathbf{x})$ is a function of the kind occuring in Taylor-type formulae, i.e. $\varepsilon|\mathbf{x}|^r$ for some $\varepsilon \in \mathbb{R}$ and $r \in \mathbf{N}$. Also, in most application $\{\mathcal{C}_i\}_{i\in\mathbb{I}}$ is a finite set. Under these very general circumstances we can conclide that $\mathcal{H}(\varepsilon|\mathbf{x}|^r)\cap\mathcal{C}$ is a convex set in $\mathcal{H} \cong \mathbb{R}^q$. Furthermore, if none of the conditions $\mathcal{C}_i$ are trivial (e.g. $I = [a, a]$ in $\mathcal{C}$-Range(I), or $c = 0$ in $\mathcal{C}$-Change(c)), then $\mathcal{H}\cap\mathcal{C}$ is a bounded convex body whose boundary is a semialgebraic set. (Such bodies are known as regular in computer-aided design.)

Theorem 6 *Let $\mathcal{H}$ be a linear subspace of $\mathcal{P}_d = \mathbb{R}[\mathbf{x}_1, \ldots, \mathbf{x}_d]$ with $\dim\mathcal{H} = q < \infty$, and let $\{\mathcal{C}_i\}_{i\in[n]}$ be a finite set of function classes of types described in Section 3. Then $A = \bigcap\limits_{i\in[n]} \mathcal{C}_i \cap \mathcal{H}$ is a convex semialgebraic set in $\mathcal{H}$. If none of the conditions defining classes $\mathcal{C}_i$ are trivial, then A is a convex body with piecewise-algebraic boundary.*

Proof. The statement that $h \in A$ can be expressed by a formula in the first-order predicate language of the reals (see Basu, Pollack, and Roy 2003). By Tarski's (1951) quantifier elimination theorem such a formula is equivalent to a formula free of quantifiers. A quantifier-free formula in the language of the reals defines a semialgebraic set. By Theorem 4 A is convex. If none of the conditions are trivial, this set has a non-empty interior. Since A is closed, convex, and has a non-empty interior, $A = \overline{\operatorname{int} A}$. Thus A is a combex gody with piecewise-algebraic boundary. ■L A is called the set of admissible functions.

5.1 Complexity of Determination of the Set of Admissible Functions

Tarski's theorem on quantifier elimination is constructive, that is, he proved that there is an algorithm that takes a a formula in the language of the reals and produces a quantifier-free equivalent formula. There is an extensive discussion of such algorithms in (Basu, Pollack, Roy 2003). The best known algorithm is by Renegar (1988). See Basu et al. (2003) for the complexity analysis in the most general case. We will only state here Renegar's bound for a special case that covers our situation.

Theorem 7 *Let Φ be a formula in the first-order predicate language of the reals. Suppose it has s sign conditions, each of degree at most r, and all quantifiers are over disjoint blocks of variables of (fixed) length b. There are ω blocks. The total number of variables is $\omega b + l$. Then the algebraic complexity of rewriting Φ into an equivalent quantifier-free form is bounded by*

$$s^{((b+1)^\omega)} r^{lO(b)^\omega}.$$

Theorem 8 *Let Φ be a formula in the first-order predicate language of the reals. Suppose it has s sign conditions, each of degree at most r, and all quantifiers are over disjoint blocks of variables of (fixed) length b. Suppose there are ω such blocks. Let $\omega b + l$ be the total number of variables*

in the formula. Then the algebraic complexity of rewriting Φ *into an equivalent quantifier-free form is bounded by*

$$s^{(b+1)^\omega} r^{lO(b)^\omega}.$$

If all the parameters in this bound are considered as variables, the bound is heavily super-exponential. However, as we will see from the following example, in any particular learning application with at most two conditions $\mathcal{C}_i$ the number of sign conditions s is a small constant (say, less than 12 and usually 3-5), the number of blocks is also small (say, less than 10). Also, in most cases the highest degree r is only 2. Furthermore, we see the complexity does not depend on the number of measurements k. If $\mathcal{C}$ is a $\mathcal{C}$-Range($[a, b]$) condition and B is a Euclidean ball, then there is only one block, i.e. $\omega = 1$ and the complexity bound takes the form of

$$3^{d+1} 2^{qO(d)}.$$

in the notation of our paper (i.e. d as in $\mathbb{R}^d$ and $q = \dim \mathcal{H}$). That is the bound is single exponential in the dimension of the space and the dimension of the space of functions $\mathcal{H}$. Since all polynomials in the sign conditions are invariant under the permutations of variables, one can try to use these symmetries to speed up the quantifier elimination. We conjecture that when B is a ball in Euclidean metric and $\mathcal{C} = \mathcal{C}$-Range($[a, b]$) is the only condition, the elimination can be done in time bounded by $const \cdot 2^q$ for some *const* (note that $d \leq q$).

5.2 Probability Class 2 Estimation

For learning applications, where one wants to estimate class 2 posterior probability, bounds for locally quadratic models in $\mathbb{R}^d$ would be desirable. Suppose $h_1, \ldots, h_q$ are the $q = 1 + d + d(d+1)/2$ monomials of degree at most 2 that form a basis of $\mathcal{P}_{2,d}$. Can we characterize the set of h's for which $h(B) \subset [a, b]$, where B is a Euclidean ball at the origin? Furthermore can we characterize the hyperplanes tangent to its boundary (where the boundary is smooth) and perform the appropriate maximization? This has applications in Markov chain Monte Carlo computation of P-values for exact tests of model validity in multifactor experiments (see Jones and ONeil, 2002).

It is of interest here to see how complicated the one dimensional quadratic case is. Inverse problems in one dimension contain some of the difficulties of higher dimensional learning problems because of the sparseness of the measurement information furnished (as in the finite Fourier moment problem of reconstructing a function from limited spectral data). We thus carry out the analysis in the inverse problem setting when the target f is approximately quadratic on an interval and known not to change by more than M between any 2 points therein, i.e. the setup is $\mathbf{x}_0 = 0$, $\mathbf{x}_0 \in B = [m, m+1] \subset \mathbb{R}^1$, and $f \in \mathcal{H}(\varepsilon(\mathbf{x})) \cap \mathcal{C}$-Change($M$) We reconstruct the value of f at 0 from data consisting of noisy integrals of f over $[m, m+1]$ which contains 0. Denote $\sum |v_j| \varepsilon(\mathbf{x}_j)$ by $R(\mathbf{v})$. Reconstruction at an arbitrary point in an arbitrary interval can be done

by a simple linear reparameterization. All integrals are over $[m, m+1]$ We have

$$f(x) = a_o + a_1(x-m)^2 + a_2(x-m) + \mathcal{E}(x), \text{ with } |\mathcal{E}(x)| < \varepsilon(x) \text{ in } [m, m+1].$$

Then

$$\mathbf{E}\left\{(F(\mathbf{v}) - f(0))^2 \,|\, \theta_1, \ldots, \theta_k\right\} =$$

$$\mathbf{v}^t\mathbf{N}\mathbf{v} + \{a_1Q_1 + a_2Q_2 + Ca_0 + v_0 + \int \sum v_j\theta_j(t)\mathcal{E}(t)dt\}^2,$$

where $C = \sum v_j \int \theta_j(t)dt - 1$, $Q_1 = \sum v_j \int (t-m)^2\theta_j(t)dt - m^2$ and $Q_2 = \sum v_j \int (t-m)\theta_j(t)dt + m$. Since a_0 is arbitrary C must be 0 for the minimax weights.

To get the minimax value in the exact quadratic case we determine the convex set in (a_1, a_2) subspace of $\mathcal{H}$ for which $u(x) = a_1x^2 + a_2x$ has change bounded by M on the unit interval. This involves solving the simultaneous inequalities:

$$|a_1c_i^2 + a_2c_i - a_1c_k^2 - a_2c_k| < M$$

for all pairs (i, k), where c_i, c_k are 2 of the (2 or 3) critical and end points for $u(x)$ in $[0, 1]$. In the case of approximately quadratic h we get the bound by using the same set with M replaced by $M + 2\mu$ where μ is the maximum of $\varepsilon(x)$ on B.

We continue the analysis for the exact quadratic case . In Fig. 1 the set A of admissible (a_1, a_2) is displayed with 6 boundary curves with end points labelled by their coordinates. The expressions for the curves as functions of a_1 are also included. This set is symmetric about the origin in the (a_1, a_2) space. So, in maximizing $|a_1Q_1 + a_2Q_2 + v_0|$ we can choose $a_1Q_1 + a_2Q_2$ to have the same sign as v_0. Hence $v_0 = 0$ will minimize the bound for any $\mathbf{v}$. So, we need only to maximize $|a_1Q_1 + a_2Q_2|$. For this we determine a tangent hyperplane to the boundary of the form $a_1Q_1 + a_2Q_2 = \pm t$ and use $\mathbf{v}^t\mathbf{v} + |t|^2$ for the best bound for fixed $\mathbf{v}$. (In the approximate case we add $\mathcal{E}(\mathbf{v})$ to $|t|$ before squaring.) If we traverse the boundary clockwise we calculate slopes between corners as functions of a_1 :

$(-4M, 4M)$ to$(-M, 2M)$ slope $=-(M/-a1)^1/2$

$(-M, 2M)$ to $(M, 0)$ slope $= -1$

$(M, 0)$ to $(4M, -4M)$ slope $= (M/a_1)^{1/2} - 2$

$(4M, -4M)$ to $(M, -2M)$ slope $= -(M/a_1)^{1/2}$

$(M, -2M)$ to $(-M, 0)$ slope $= -1$

$(-M, 0)$ to $(-4M, 4M)$ slope $= (M/-a_1)^{1/2} - 2$

Given Q_1, Q_2 we know the slope and hence we can identify either a point of tangency to one of the curves or at one of the corners. From this we can identify $\pm t$ and therefore $|t|$.

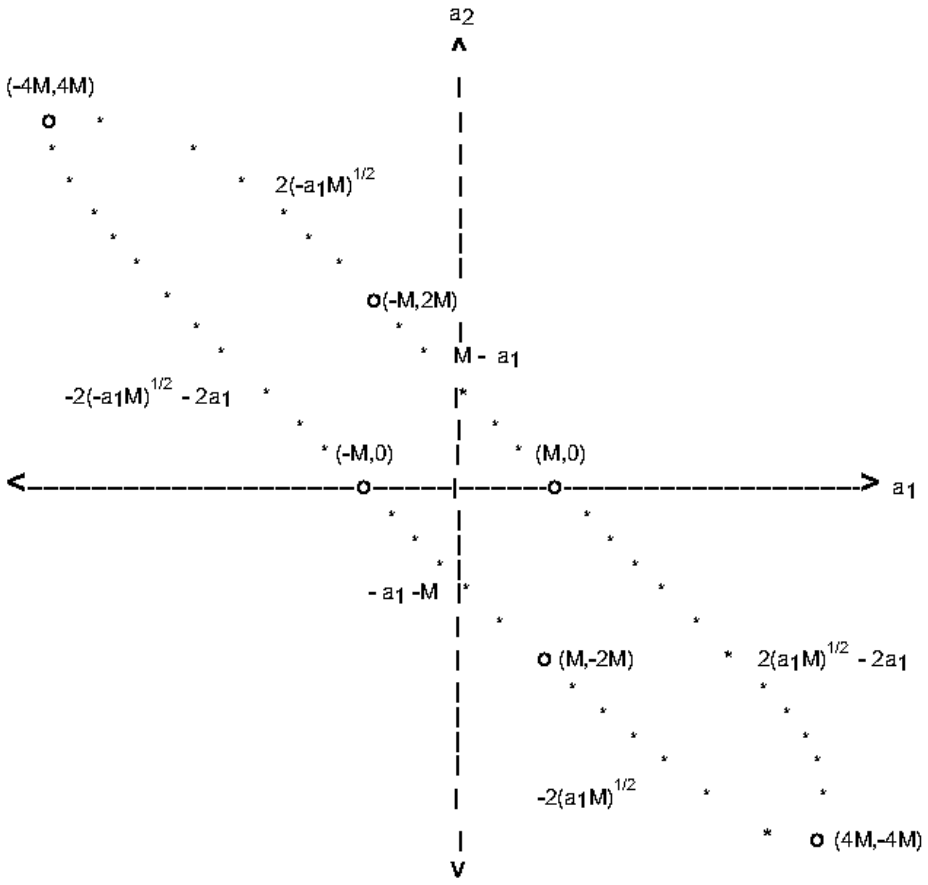

Figure 1: Boundary of the Section $a_0 = 0$ of the Set of Admissible Functions

So, by a straightforward but extremely tedious calculation we find that the maximum of $|a_1Q_1 + a_2Q_2|$ is given by $U(M, Q_1, Q_2) =$

$|MQ_2|$ if $Q_1 = Q_2$

$|M\frac{Q_2^2}{Q_1}|$ if $1/2 < \frac{Q_1}{Q_2} < 1$

$\frac{|2MQ_2 - MQ_1|}{(2-(Q_1/Q_2))^2}$ if $1 < \frac{Q_1}{Q_2} < \frac{3}{2}$

$4M|Q_1 - Q_2|$ otherwise.

For the upper bound on the minimax value we use $U(M + 2\mu, Q_1, Q_2) + R(\mathbf{v})$. We restate this as

Theorem 9 *Let $\mathcal{H} = \mathcal{P}_{2,1}$ be the space of quadratic functions of one variable. parametric family be given by $f(x; a) = a_0 + a_1(x - m)2 + a_2(x - m)$. Assume that in $B = [m, m + 1]$ the unknown function $f(x)$ is within $\varepsilon(x)$ of some member of $\mathcal{H}$. Let the loss function be the squared error one. Suppose $f \in \mathcal{C}$-Change$([m, m + 1])$. Let*

$$\mathcal{L}(\mathbf{v}) = \mathbf{v}^t\mathbf{S}\mathbf{v} + \{U(M + 2\mu, Q_1, Q_2) + \int_{t\in\mathbb{R}^d} \sum |v_j\theta_j(t)|\varepsilon(t)dt\}^2$$

and let

$$\mathcal{I}(\mathbf{v}) = \mathbf{v}^t\mathbf{S}\mathbf{v} + \{U(M, Q_1, Q_2)\}^2$$

Then the mean squared error of an estimator $F(\mathbf{v})$ is bounded from above by $\mathcal{L}(\mathbf{v})$ and the mean squared error of the optimal estimator F in the exact quadratic case is bounded from above by $\min_{\mathbf{v}\in\mathbb{R}^3} \mathcal{I}(\mathbf{v})$.

6 Complexity of Minimax Algorithm

The complexity of determination of the set of admissible functions $A = \bigcap_{i\in\mathbb{I}} \mathcal{C}_i \cap \mathcal{H}$ is estimated in Section 5. What is the complexity of finding

$$\min_F \max_{h\in\mathcal{H}(\varepsilon|\mathbf{x}|^r)\cap\mathcal{C}} \mathbf{E}_\mathsf{P}\{L(F, h) \mid \mathbf{x}_1, \ldots, \mathbf{x}_k\}$$

and

$$\arg\min_F \max_{h\in\mathcal{H}(\varepsilon|\mathbf{x}|^r)\cap\mathcal{C}} \mathbf{E}_\mathsf{P}\{L(F, h) \mid \mathbf{x}_1, \ldots, \mathbf{x}_k\}$$

for the l_p loss function, i.e. for the loss function of the form $|\ |^p$?

For the l_p loss function the minimax choice of F corresponds to the value of $\mathbf{v}$ for which the null-subspace of $\mathbb{R}^q$ is equidistant from the maximum and minimum hyperplanes in the direction specified by $\mathbf{x}_0$; for example, if $\mathcal{H} = \mathcal{P}_{r,d}$ the direction is $(1, x_1, \ldots, x_d, x_1, x_2, \ldots...)$, the vector of all monomials .of degrees 0 through r in the coordinates of $\mathbf{x}_0 = (x_1, \ldots, x_d)$. The most diffcult part here is finding the maximal and minimal values of the expected loss in the direction determined by $\mathbf{x}_0$. This ia classical optimization problem with convex feasibility set and linear objective function. Furthermore, we have the following conjecture.

Conjecture 10 *Let $\mathcal{H}$ be a linear subspace of $\mathcal{P}_d = \mathbb{R}[\mathbf{x}_1, \ldots, \mathbf{x}_d]$ with $\dim \mathcal{H} = q < \infty$. Suppose $B \subset \mathbb{R}^d$ is a ball in l_p metric, and let $\{\mathcal{C}_i\}_{i\in[n]}$ be a finite list of function classes, defined by nontrivial conditions of types described in Section 3. Then the membership predicate for $A = \bigcap_{i\in\mathbb{I}} \mathcal{C}_i \cap \mathcal{H}$ can be writtenas a conjunction of at most $(2n+1)^{d+1}2^{qO(d)}$ sign conditions are of affine or quadratic type. Thus, A is a convex body defined by convex quadratic inequalities.*

The analysis of this conjecture will be included in a subsequent journal paper. The significance of this theorem is in that for quadratically constrained linear programming there are fast interior-point algorithms, which are efficient from both practical and theoretical point of view. Three major families of such algorithms are the Potential Reduction and Karmarkar Methods. Two most well-studied Potential Reduction methods and the Path Following and Primal-Dual methods. For any given precision ε they all compute an ε-approximation of the argument of the objective function in time (in the algebraic model of computation), which is linear in $\log_2 \varepsilon^{-1}$ and polynomial in the number of constraints m and variables q. The bit complexity of these methods is linear in $\log_2 \varepsilon^{-1}$ and polynomial in the number of constraints m, variables q, and the maximum L of the binary sizes of the numbers in the input data. The number of interior point steps for some of these methods is bounded by $O(m)$ and for some $O(\sqrt{m})$ The best theoretical bound, given by the path following and primal-dual methods has the complexity of $O\left(m^1 2(m+n)n^2 L \log L \right)$ per each accurate digit of the approximation. For details see Nemirovskii (2004).

References

[1] Basu S., Pollack R., Roy M.-F. (1996), On the Combinatorial and Algebraic Complexity of Quantifier Elimination, *Journal of the ACM* , Nov. 1996, **Vol. 43**, No. 6, pp. 1002-1046.

[2] Basu S., Pollack R., Roy M.-F. (2003) *Algorithms in Real Algebraic Geometry*, Springer-Verlag.

[3] Jones, L.K. (2006) *Optimal Local Statistical Learning of Functions with Best Finite Sample Estimation Error Bounds: Applications to Ridge and Lasso Regression, Boosting, Tree Learning, Kernel Machines and Inverse Problems.* Submitted. Prebulication version is available as Tech. Rep. 2006-003, Department of Computer Science, UMASS Lowell, `http://teaching.cs.uml.edu/~heines/techrpts/`.

[4] Jones, L.K. and ONeil, P.J. (2002) Markov Chain Monte Carlo Algorithms for Computing Conditional Expectations Based on Sufficient Statistics, *JCGS*, **vol. 11**, No. 3, Sept.2002, pp. 660-677.

[5] Nemirovski A. (2004), *Interior Point Polynomial Time Methods in Convex Programming. Lecture Notes.*, Georgia Institute of Technology, School of Industrial and Systems Engineering, Technical Report ISYE 8813.

[6] Poggio,T. and Smale,S. (2003) The Mathematics of Learning: Dealing with Data, Notices of the AMS, **Vol. 50**, No. 5, May 2003, 537-544

[7] Renegar J.(1992) On the Computational Complexity and Geometry of the First-order Theory of the Reals. Part III: Quantifier Elimination, *J. SYMBOLIC COMP.* , **Vol. 13**, March 1992, No. 3, pp. 329-352.

[8] Tarski A. (1951) *A Decision Method for Elementary Algebra and Geometry.* Prepared for publication by J.C.C. McKinsey. University of California Press.